



University  
of Glasgow

Hu, Yi-Ming (2015) *Novel inference methods for gravitational wave astrophysics*. PhD thesis.

<http://theses.gla.ac.uk/6441/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

# NOVEL INFERENCE METHODS FOR GRAVITATIONAL WAVE ASTROPHYSICS

YIMING HU

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF  
*Doctor of Philosophy*

SCHOOL OF PHYSICS AND ASTRONOMY  
COLLEGE OF SCIENCE AND ENGINEERING  
UNIVERSITY OF GLASGOW

MAY 14, 2015

© YIMING HU

# Preface

This thesis based on the work of the author, Yi-Ming Hu, during my stay at the University of Glasgow. From October 2011 to March 2015, under the joint supervision of Prof. Martin Hendry and Dr. Ik Siong Heng. We collaborated on a number of projects sharing the same theme of data analysis methodology related to Gravitational Waves.

In Chapter 1, I introduce basic concepts of gravitational waves, their generation and detection. A brief history of Gravitational Wave experiments is introduced, and the anticipated sources are listed and discussed. In addition, I discuss the role of data analysis in making detections of Gravitational Waves, and what new science can we expect to make with the help of Gravitational waves. The majority of this chapter is presented based on published literature.

In Chapter 2, I discuss the power of Bayesian Inference and how to apply it to Gravitational Wave data analysis. For the anticipated compact binary coalescence signals, the optimal search method of matched filtering is introduced. The principles of Bayesian Inference are also introduced, including the basic concepts, their difference from a frequentist approach and the important potential for performing Bayesian Model Selection using the Bayes Factor. A brief review of existing stochastic sampling strategies is presented; this includes a detailed introduction to the algorithms of Markov Chain Monte Carlo and Nested Sampling, as well as numerous implementations of these algorithms. The majority of this chapter is summarised from published literature.

In Chapter 3, Bayesian Model Selection is performed on an astrophysical problem of distinguishing which model is better supported by the observed timing data for the magnetar 1E 2259+586 – either the ‘successive anti-glitches’ model or the ‘anti-/normal glitch pair’ model. A simple description of the parametric model of pulsar timing is made, and a Nested Sampling algorithm is applied to the timing data, based on both successive anti-glitches and

an anti-/normal glitch pair. This work was proposed by me and carried out mainly by myself, with helpful discussions and input from Matthew Pitkin, Ik Siong Heng and Martin Hendry.

In Chapter 4, a novel MCMC algorithm designed for efficient sampling of multi-modal posterior, known as *mixed MCMC* is proposed. I discuss the motivation and implementation of this novel algorithm, including an application to a toy model and a comparison of the results with a theoretical calculation. This algorithm was initially proposed by me, and the work was mainly carried out by myself.

In Chapter 5, I discuss the global optimisation of future generations of Gravitational Wave detectors. I introduce the Figures of Merit used to judge the scientific output and effectiveness of a Gravitational Wave detector network. The algorithm of mixed MCMC is then applied to sample efficiently the high dimensional parameter space that represents the detector network. The optimisation is performed for both a 3-detector-network and a 5-detector-network. This work was carried out in close collaboration with colleagues from several institutions, with the majority of the work designing the methodology and implementing it carried out by myself.

In Chapter 6, the setup and analysis of the significance Mock Data Challenge, also known as the “Hamlet Test”, is described. In this Mock Data Challenge, two different types of method for estimating the significance of Gravitational Wave triggers are compared against the theoretical significance. These methods are, when estimating the background distribution, either to remove the foreground events or to keep them in. We discuss results for simple, realistic and complicated background distributions, as well as for zero, low, medium and high foreground event rates. This work was carried out in close collaboration with colleagues from multiple institutions, with the majority of Mock Data generation and the comparison, analysis and interpretation of the results done by myself.



# Summary

With the development of more and more elegant and sensitive interferometric gravitational wave detectors, we are expecting the first direct detection of gravitational waves in a short time. This triggers huge interest to develop more powerful tools to perform data analysis on these signals, and to develop a good understanding of the analysis so that confident conclusions can be made. A further step would be to view into the future, as the first detections will boost the scientific demands for more powerful future generation detectors, which identifies the task of optimising the site of such detectors.

Bayesian Inference plays a vital role in data analysis, and one excellent example that demonstrates its usefulness is its ability to resolve the tension between multiple models using the methodology of Bayesian Model Selection. In this thesis we apply this methodology to the timing data of pulses from the pulsar 1E 2259+586. With a set of different choices for the prior range, a fair and quantitative comparison can be made between two competing models: that of so-called successive anti-glitches and an anti-/normal glitch pair. Our analysis of the data shows a consistent support for the successive anti-glitches model, with a Bayes Factor of  $\sim 45$ , where the uncertainty has been estimated from nested sampling and from multiple runs that are slightly different, but still within a factor of two, showing a general consistency. Simplifying the timing model will only make the Bayes Factor even bigger, while the two event model is overwhelmingly supported over the one event model.

In gravitational wave data analysis, posteriors are generally complicated structures containing multiple modes. A novel algorithm to achieve efficient sampling for multi-modal posteriors, known as mixed MCMC, is proposed in this thesis. This enables communication between multiple regions within the parameter space by adopting a novel jump proposal. We present the mixed MCMC algorithm and first apply it to a toy model problem, where the likelihood may be determined theoretically. By comparing the theoretical and empirically sampled values of  $2 \log(\mathcal{L})$  for credible regions that correspond to 68.27%, 95.45% and

99.73%, we conclude that for our illustrative model the sampling result of mixed MCMC is consistent with the theoretical prediction with small uncertainty. Since it does not require multiple chains with different temperatures, mixed MCMC can boost the efficiency of sampling by design, compared with (for example) parallel tempering MCMC.

The sampling strategy of mixed MCMC can be helpful for not only Bayesian Inference, but also more general problems like the global optimisation of future generations of Gravitational Wave Detectors. As we expect such problem to be intrinsically high dimensional and multi-modal, mixed MCMC is a suitable sampling method, and we develop and apply it in this thesis. Based on our analysis it is concluded that for both a 3-detector-network and a 5-detector-network, Australia hosts the “best” site, in the sense that such site is most flexible, i.e. it can be involved in the largest number of detector networks, involving different component sites, that have a high ‘Figure of Merit’.

The work of gravitational wave data analysis leads to the ultimate goal of making a direct detection of gravitational waves, which in turn requires the ability of distinguish astronomical signals from a noisy background, and assess the significance of each gravitational wave ‘trigger’ (i.e. candidate event) appropriately. There are two types of method for estimating significance and these differ by the key distinction of either removing the foreground events from the background estimation or keeping them in the analysis. This thesis presents the results of a Mock Data Challenge (MDC), carried out within the LIGO Scientific Collaboration using different data analysis pipelines, designed to investigate these two methods for estimating significance. It contains a variety of background complexity ranging from simple, realistic to complex, and foreground event rate ranging from zero, low, medium and high. Analysis of the MDC results illustrated that generally all methods for determining the significance agree well with each other, irrespective of the background complexity. However, a discrepancy became apparent between the results for removal or non-removal of foreground events, for events below a significance level of  $< 10^{-3}$ . Our results demonstrated that the removal method is an unbiased estimator for the mean of the significance. However, as the most scientifically interesting events are likely to have a very small numerical value for their significance, such method would overestimate that significance for most of the realisations.

# Acknowledgements

I would like to thank my supervisors Martin Hendry and Ik Siong Heng, who are both excellent mentors and good friends, from whom I learnt much more than researching. You allow me to follow my interests and to try new things, and you gave me priceless suggestions, I deeply appreciate how lucky I am.

I would also like to give my sincere thanks to my many collaborators especially Szabolcs Márka, Chris Messenger, Matthew Pitkin, it's always inspiring to discuss with you, and I thank you for being so patient, so positive and encouraging. And I want to thank Xi-Long Fan, although we always disagree in many aspects, but we can always find agreements in the passion of the gravitational wave. I'm grateful for the experience of collaboration with Péter Raffai, László Gondán, Nándor Kelecsényi, Zsuzsa Márka, Collin Capano, Thomas Dent, Chad Hanna, John Veitch and Matt West, the brain storm with you are pleasant memories. I won't be too sentimental here as I anticipate to have more collaborations with you in the longer term.

I appreciate the China Scholarship Council (CSC) and the University of Glasgow for funding me in the passing years. Without such financial support it would be impossible for me to pursue my academic path. I would also want to thank the Principals Early Career Mobility Fund to sponsor my visit to Columbia University in the City of New York, to perform research and to make good friends.

Many thanks to Craig Lawrie, Ignacio Santiago, Erin MacDonald, Josh Logue, Gareth Davies, Jade Powell, Brynley Pearlstone, Man-Leong Chan, Jessica Steinlechner, Sheena Barclay and so many other nice people in the Institute for Gravitational Research and the School of Physics and Astronomy, the name list can go on and on, but my dear friends, I won't forget the days we share laugh and whiskey in the beautiful Scotland. You really make Glasgow such a pleasant place for me to memorise.

I can't be more grateful for the support from my family and my girlfriend. My dear sister, Wei Hu, you opened the gate of science to me, inspired my curiosity to the Nature and especially to the starry night. My dear parents, Yong-Xiang Hu and Ya-Ping Zhu, I know you didn't particularly want me to study such a "useless" subject, but nevertheless you bought me a telescope at a financially tough time. When I determined to have astronomy as my major, you had no objection any longer, but only support wholeheartedly. My dear girlfriend, De-Yue Yang, my sincere thank to you for accompanying me for 7 years, following me to this country. You are the stars of my life.

“道天制志天道  
学文言研文学”

“To discover the celestial principles and devote to the moral law;  
To learn the classics and to study the literature.”

— a Chinese acrostic I wrote at the age of 10,  
which contains the word “astronomy (天文)” twice.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Gravitational Wave formalism . . . . .	3
1.2	Gravitational Wave from Compact Binary Coalescence . . . . .	5
1.3	Detecting Gravitational Wave with Ground-Based Detectors . . . . .	7
1.4	A brief history of Gravitational Wave experiment . . . . .	9
1.5	Astronomical Sources for Ground-based Gravitational Wave Detectors . . .	15
1.6	The Role of Data Analysis towards Detecting Gravitational Waves . . . . .	19
1.7	Physics and Astronomy with Gravitational Waves . . . . .	20
<b>2</b>	<b>Applying Bayesian Inference to Gravitational Wave Data</b>	<b>24</b>
2.1	Matched Filtering . . . . .	25
2.2	Bayesian Inference . . . . .	27
2.3	The Bayes Factor and Model Selection . . . . .	29
2.4	Stochastic Samplers . . . . .	32
2.4.1	Markov Chain Monte Carlo (MCMC) . . . . .	32
2.4.2	Nested Sampling . . . . .	35
2.4.3	Parallel Tempering MCMC . . . . .	37
2.4.4	Delayed Rejection MCMC . . . . .	38
2.4.5	Reversible Jump MCMC . . . . .	39
2.4.6	Automated MCMC . . . . .	39
2.4.7	Affine Invariant MCMC . . . . .	40
2.4.8	MultiNest . . . . .	40
2.4.9	Combinations . . . . .	41

<b>3</b>	<b>Bayesian Model Selection on Anti-Glitch Models</b>	<b>42</b>
3.1	Timing Model . . . . .	43
3.2	Results . . . . .	46
3.2.1	Setting the Priors . . . . .	46
3.2.2	Comparing the Models . . . . .	49
3.2.3	Robustness Check . . . . .	49
3.3	Discussion and Conclusions . . . . .	53
<b>4</b>	<b>Mixed MCMC</b>	<b>55</b>
4.1	Method . . . . .	56
4.1.1	Markov Chain Monte Carlo . . . . .	56
4.2	mixed MCMC . . . . .	57
4.2.1	Algorithm . . . . .	58
4.3	Toy Model . . . . .	60
4.3.1	Analytical Evaluation of $2\Delta \log(\mathcal{L})$ . . . . .	61
4.3.2	Application of the mixed MCMC . . . . .	63
4.4	Discussion . . . . .	64
<b>5</b>	<b>Global Optimisation for Future Gravitational Wave Detectors Sites</b>	<b>67</b>
5.1	Figures of Merit for GW Detector Networks . . . . .	70
5.2	Description of Method . . . . .	72
5.2.1	Methodology . . . . .	73
5.2.2	The actual realisation . . . . .	75
5.2.3	Constructing the partition . . . . .	76
5.2.4	Exclusion of unsuitable regions . . . . .	77
5.2.5	Optimising parameters . . . . .	79
5.3	Results . . . . .	83
5.3.1	Network of 3 detectors . . . . .	85
5.3.2	Network of 5 detectors . . . . .	88
5.4	Conclusions and Discussion . . . . .	90

<b>6</b>	<b>The Significance Mock Data Challenge</b>	<b>94</b>
6.0.1	Background of the Mock Data Challenge . . . . .	95
6.1	The mock data challenge . . . . .	99
6.1.1	Modelling the detector background . . . . .	101
6.1.2	Modeling an astrophysical foreground . . . . .	102
6.1.3	The definition of FAP for the MDC . . . . .	103
6.2	Background estimation algorithms . . . . .	104
6.2.1	ihope: FAP via inverse false alarm rate (IFAR) . . . . .	104
6.2.2	The gstlal approach . . . . .	107
6.2.3	A new approach . . . . .	107
6.3	Results . . . . .	110
6.3.1	Direct comparison with exact FAP . . . . .	110
6.3.2	Self consistency tests: $p$ - $p$ plots . . . . .	118
6.3.3	Receiver Operating Characteristic (ROC) analyses . . . . .	121
6.3.4	Boxplots . . . . .	126
6.3.5	Uncertainty in estimation . . . . .	133
6.4	Discussion . . . . .	135
<b>7</b>	<b>Conclusion and Future work</b>	<b>138</b>
<b>A</b>	<b>Additional MDC results</b>	<b>140</b>
A.1	Cumulative Distribution Function of SNR . . . . .	140
A.2	Direct comparison . . . . .	147
A.3	Receiver Operating Characteristic plots . . . . .	158
<b>B</b>	<b>Parameters for trigger distribution</b>	<b>162</b>
	<b>Bibliography</b>	<b>166</b>



# List of Tables

1.1	Estimated detection rates for the initial LIGO and the advanced LIGO detector era for different sources. . . . .	17
2.1	odds ratio value and their corresponding strength of evidence . . . . .	32
3.1	Parameter estimation for the timing model from TEMPO2 fitting . . . . .	47
3.2	Upper and lower boundaries for $n = 3$ . . . . .	48
3.3	Upper and lower boundaries for $n = 5$ . . . . .	48
3.4	Upper and lower boundaries for $n = 10$ . . . . .	48
3.5	Estimated Bayes Factors with uncertainties for different $n$ values . . . . .	52
3.6	Bayes Factors of different realisations with mean and standard deviation. . .	53
4.1	Likelihoods corresponding to 68.27%, 95.45% and 99.73% credible regions with equal $\sigma$ . . . . .	63
4.2	Likelihoods corresponding to 68.27%, 95.45% and 99.73% credible regions with $\frac{\sigma_2^x \sigma_2^y}{\sigma_1^x \sigma_1^y} = 0.5$ . . . . .	64
5.1	List of sites for component detectors of networks containing the global optimal position, for a 3 detector network. . . . .	90
5.2	List of sites for component detectors of networks containing the global optimal position, for a 5 detector network. . . . .	91
6.1	Classification of all 14 experiments in background complexity and foreground rate . . . . .	101
6.2	Detection efficiency (in percent) at given false positive rate, together with the uncertainty considering a binomial likelihood. . . . .	127
B.1	parameters for background distribution . . . . .	163

B.2	Parameters for background distribution (tail). . . . .	164
B.3	parameters for rates of both background and foreground signals . . . . .	165

# List of Figures

1.1	Orbital decay of PSR B1913+16 . . . . .	2
1.2	Illustration of the two gravitational wave polarisations. . . . .	5
1.3	Illustration of the principle of laser interferometer gravitational wave observatory. . . . .	8
1.4	Illustration of the definition of the two directional angles and the polarisation angle. . . . .	10
1.5	Bar detector in Munich . . . . .	11
1.6	The power spectral density for the bar detector Auriga. . . . .	12
1.7	Sensitive volume for initial LIGO and advanced LIGO. . . . .	13
1.8	Expected noise budget of the advanced LIGO . . . . .	14
1.9	Measured value of the gravitational constant $G$ from different research groups. . . . .	21
2.1	An illustration of likelihood contours and corresponding $X(L)$ . . . . .	36
3.1	Observational timing data of 1E 2259+586 together with the X-ray flux. . . . .	44
3.2	Posterior contours for the second anti-glitch event. . . . .	50
3.3	Timing residual of two successive anti-glitch events model versus one anti-glitch event model. . . . .	51
4.1	Example realisation of mixed MCMC applied to the toy model posterior. . . . .	65
5.1	Division of the world map into 6 distinct sections. . . . .	76
5.2	World map with various of exclusion criteria. . . . .	78
5.3	Relative change of the $I$ FoM, averaged over all possible combinations from the hypothetical Earth, as a function of <i>healpix</i> input parameter. . . . .	80
5.4	Average number of samples for a 3-detector network using different numbers of CPUs. . . . .	82

5.5	Individual normalised histogram of $I$ , $D$ and $R$ . . . . .	84
5.6	World map of flexibility index for a 3-detector gravitational wave detector network. . . . .	86
5.7	Zoom in from figure 5.6a, maps of flexibility index for a 3-detector gravitational wave detector network. . . . .	87
5.8	Zoom in from figure 5.6b, maps of flexibility index for a 3-detector gravitational wave detector network. . . . .	88
5.9	World map of flexibility index for a 5-detector gravitational wave detector network. . . . .	89
5.10	Zoom in from figure 5.9b, maps of flexibility index for a 5-detector gravitational wave detector network. . . . .	91
6.1	Cartoon illustration of time slide method. . . . .	97
6.2	Direct comparisons of FAP estimates with the true FAP for experiment 14. .	112
6.3	Direct comparisons of FAP for experiment 9. . . . .	114
6.4	Reverse cumulative distribution function (CDF) of the triggers' SNR for experiment 6 and 9. . . . .	115
6.5	Direct comparisons of FAP for experiment 6. . . . .	116
6.6	Direct comparisons of FAP for experiment 11. . . . .	117
6.7	$p$ - $p$ plot for experiment 1. . . . .	119
6.8	$p$ - $p$ plot for experiment 3. . . . .	120
6.9	$p$ - $p$ plot for experiment 12. . . . .	120
6.10	$p$ - $p$ plot for experiment 14. . . . .	121
6.11	ROC plot for experiment 2. . . . .	123
6.12	ROC plot for experiment 8. . . . .	124
6.13	ROC plot for experiment 4. . . . .	125
6.14	Boxplot for log of ratio between estimation and exact, drawn on data with FAP between $10^{-3}$ and $10^{-2}$ . Left means optimistic while right is conservative. The upper 3 are with high rate and the bottom 4 are with zero rate. . .	129
6.15	Boxplot for log of ratio between estimation and exact, drawn on data with FAP between $10^{-4}$ and $10^{-3}$ . Left means optimistic while right is conservative. The upper 3 are with high rate and the bottom 4 are with zero rate. . .	131

6.16	Boxplot for log of ratio between estimation and exact, drawn on data with FAP between $10^{-5}$ and $10^{-4}$ . Left means optimistic while right is conservative. The upper 3 are with high rate and the bottom 4 are with zero rate. . .	134
A.1	Reverse CDF distribution of the triggers' SNR for experiment 1. . . . .	141
A.2	Reverse CDF of the triggers' SNR for experiment 2. . . . .	141
A.3	Reverse CDF of the triggers' SNR for experiment 3. . . . .	142
A.4	Reverse CDF of the triggers' SNR for experiment 4. . . . .	142
A.5	Reverse CDF of the triggers' SNR for experiment 5. . . . .	143
A.6	Reverse CDF of the triggers' SNR for experiment 7. . . . .	143
A.7	Reverse CDF of the triggers' SNR for experiment 8. . . . .	144
A.8	Reverse CDF of the triggers' SNR for experiment 10. . . . .	144
A.9	Reverse CDF of the triggers' SNR for experiment 11. . . . .	145
A.10	Reverse CDF of the triggers' SNR for experiment 12. . . . .	145
A.11	Reverse CDF of the triggers' SNR for experiment 13. . . . .	146
A.12	Reverse CDF of the triggers' SNR for experiment 14. . . . .	146
A.13	Direct comparisons on experiment 1. . . . .	148
A.14	Direct comparisons on experiment 2. . . . .	149
A.15	Direct comparisons on experiment 3. . . . .	150
A.16	Direct comparisons on experiment 4. . . . .	151
A.17	Direct comparisons on experiment 5. . . . .	152
A.18	Direct comparisons on experiment 7. . . . .	153
A.19	Direct comparisons on experiment 8. . . . .	154
A.20	Direct comparisons on experiment 10. . . . .	155
A.21	Direct comparisons on experiment 12. . . . .	156
A.22	Direct comparisons on experiment 13. . . . .	157
A.23	ROC plot for experiment 5. . . . .	158
A.24	ROC plot for experiment 6. . . . .	159
A.25	ROC plot for experiment 7. . . . .	159
A.26	ROC plot for experiment 9. . . . .	160
A.27	ROC plot for experiment 10. . . . .	160

A.28 ROC plot for experiment 11. . . . .	161
A.29 ROC plot for experiment 13. . . . .	161

# Chapter 1

## Introduction

Gravitational Wave (GW) science is a very exciting field for astronomy and more generally for physics. The existence of such phenomena was predicted by Einstein in the year 1916 [1, 2, 3, 4] and yet it remains undetected directly. The indirect detection of GWs was initially made by [5] in 1982, and further followed by [6], which observed the orbital evolution of the PSR B1913+16 (Hulse-Taylor binary pulsar) [7, 8]. The measured orbital evolution of this system is consistent with the gravitational radiation predicted by General Relativity (GR) to within 0.2% [6]. In figure 1.1, the observation and the theoretical prediction for the orbital evolution is depicted, which shows excellent consistency. In another binary pulsar system PSR J0737-3039 a stronger constraint of GR was obtained, showing the observation and theoretical prediction from GR agree to within a 0.05% uncertainty [9]. This confirmed that Einstein's theory of gravity, General Relativity (GR), provided a correct description of gravitational phenomena to a very high accuracy.

A direct detection of GW has become the holy grail in gravitational research, but still we have not yet reached the experimental sensitivity required to achieve this goal. In the year of 2014, the BICEP2 group announced a detection of  $B$ -mode polarisation in the Cosmic Microwave Background (CMB) [10]. At first they interpreted this result as the first evidence for inflationary gravitational waves (IGWs), but now it is generally believed that their observations demonstrated the effect of polarised Galactic dust emission [10]. Even if we could detect the  $B$ -mode polarisation in the CMB, that would still be indirect evidence for the existence of GW.

Among the four fundamental forces, gravity is the least understood one. Ever since the publication of the theory of General Relativity almost a century ago, enormous effort has been spent attempting to understand the nature of gravity, with numerous alternative gravity theories being proposed. All such theories, as long as they do not allow 'action at a distance', will intrinsically imply the existence of gravitational waves and the theories only differ by the form of GW predicted in each theory. GR is, however, still among the most promising

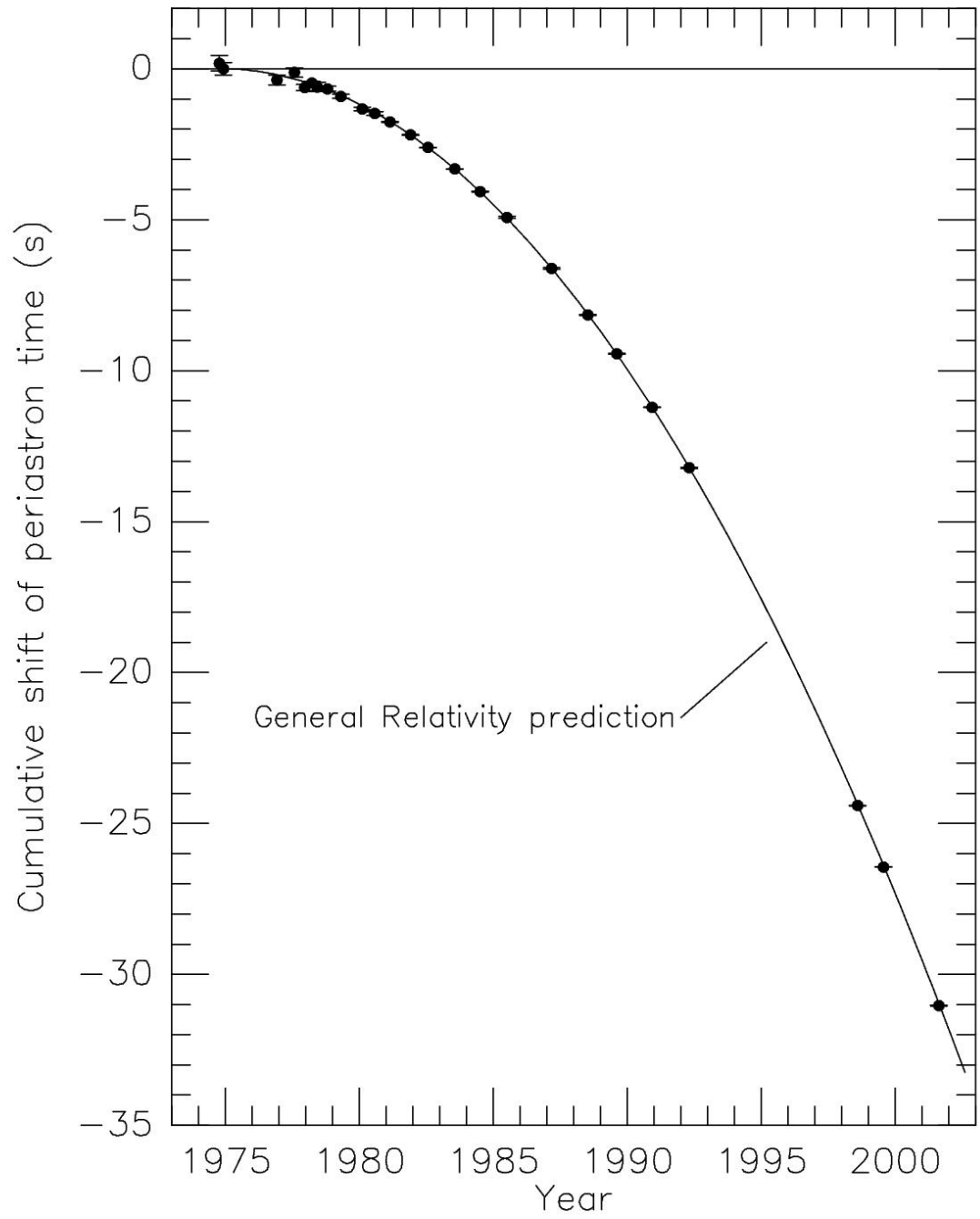


Figure 1.1: Orbital decay from the binary pulsar system PSR B1913+16. The points are actual observations with uncertainty, while the line is predicted by GR. Reproduced from [6]



theories, as no evidence of deviation from GR has been so far observed [11].

GWs are the only observational channel that we can envisage with which to explore directly the physics of strong gravitational fields. All other channels have to rely on some assumptions or theories and thus require an indirect link to be made in order to extract information about the fields. We are therefore expecting to gain a huge amount of new physics and astronomy knowledge by looking at the GW waveforms directly. For example, the detailed mechanism of supernovae is still under debate, but directly observing the gravitational wave information from core collapse supernovae will greatly help deepen our understanding of the dynamical mechanism that describes them [12]. By carefully studying the GW waveforms from binary Neutron Star (BNS), we can study the Equation of State (EoS) of the neutron material in high pressure, which can never be studied anywhere else. The combination of information from GW and other channels, known as multi-messenger astronomy, can also help understanding of the origin of short Gamma Ray Bursts (sGRBs) [13, 14]. Multi-messenger astronomy can also be viewed as an independent method of studying cosmology, providing a cross check for the cosmological parameters [15]. However, all of the above scientific targets require a huge amount of study beforehand, in which we would need to carry out theoretical study as well as computationally-intensive data analysis, to unveil such scientific treasures.

## 1.1 Gravitational Wave formalism

The interesting GW sources generally originate far away from the Earth. Combined with the fact that dipole radiation is not allowed according to GR (see below), and the coupling between GW and normal matter is tiny [1, 4, 3, 2], the actual GW signal on Earth is generally very weak. With various noise sources, we expect the first detected GW signal to be relatively weak and embedded in noise. In order to enhance the chance of detection, one needs to obtain as much knowledge as possible about the GWs, especially for sources with highly predictable waveforms [2, 16, 17].

According to GR, gravitational radiation only occurs when the quadrupole and higher order moments of a system has changed [1, 4, 3, 2]. We can define the quadrupole of a system with density  $\rho$  as [2]

$$Q_{jk} = \int \rho x_j x_k d^3x. \quad (1.1)$$

We have to note that this equation is only valid when the gravitational field within the system is not strong [4]. Fortunately, even in situations where the internal gravity is strong, there exists an alternative method to compute the quadrupole [4]. The observed gravitational wave can be characterised by the *strain* that it produces – *i.e.* the fractional change in the separation

between two points initially separated by proper distance  $l$  – which is defined by [3]

$$h = \frac{\delta l}{l}. \quad (1.2)$$

For a system with changing quadrupole, it will produce a GW signal  $h_{jk}$  at a distance  $r$  given by [4, 2]

$$h_{jk} = \frac{2G}{r c^4} \ddot{Q}_{jk} \sim \frac{G}{c^4} \frac{E_{kin}^{nonsp}}{r}. \quad (1.3)$$

as an order of magnitude estimation, and  $E_{kin}^{nonsp}$  is the non-spherical part of the system's kinetic energy. For a far zone approximation where the distance to the source is much larger than the GW wavelength, the observer's metric can be expressed as  $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ .

In GR, there are more equations than conditions, so there exist redundant degrees of freedom in the Lagrangian [18]. Usually in terms of GR it's convenient to adopt the *transverse traceless gauge*, or *TT gauge* [3, 18, 2]. Under the TT gauge, we can write

$$h_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

or, if we define  $e_+ = e_x \otimes e_x - e_y \otimes e_y$  and  $e_\times = e_x \otimes e_y + e_y \otimes e_x$ , then we can simplify to  $h = h_+ e_+ + h_\times e_\times$ . Here  $e_x = (0 \ 1 \ 0 \ 0)^T$  and  $e_y = (0 \ 0 \ 1 \ 0)^T$  [2, 4, 3]. This illustrates that there are two polarisation modes for a gravitational wave, namely the plus-polarisation and cross-polarisation, as illustrated in figure 1.2.

It was not trivial for people to finally agree that GWs do contain energy, instead of simply a coordinate transforming effect, and furthermore, like all other forms of energy, GWs can themselves induce GWs. The power passing per unit area carried by a GW is determined by [3, 4]

$$\frac{dE}{dt dA} = T_{03}^{GW} = -\frac{c^3}{16\pi G} \langle \dot{h}_+^2 + \dot{h}_\times^2 \rangle = -\frac{G}{8\pi c^5 r^2} \langle \ddot{I}_{ij}^{TT} \ddot{I}_{TT}^{ij} \rangle \quad (1.4)$$

where  $I$  is the symmetric reduced quadrupole moment, which differs from the definition of  $Q$  by the trace, and the angle bracket means average over several wavelengths [4]. One can obtain the GW luminosity  $L_{GW}$  of a GW system by integrating the equation 1.4 [1, 4, 3]

$$L_{GW} = -\frac{dE}{dt} = \frac{G}{5c^5} \langle \ddot{I}_{ij} \ddot{I}^{ij} \rangle. \quad (1.5)$$

The pre-factor  $\frac{G}{c^5}$  indicates that the luminosity for a GW system tends to be very small, unless for extreme scenarios where  $\langle \ddot{I}_{ij} \ddot{I}_{TT}^{ij} \rangle$  is very large.

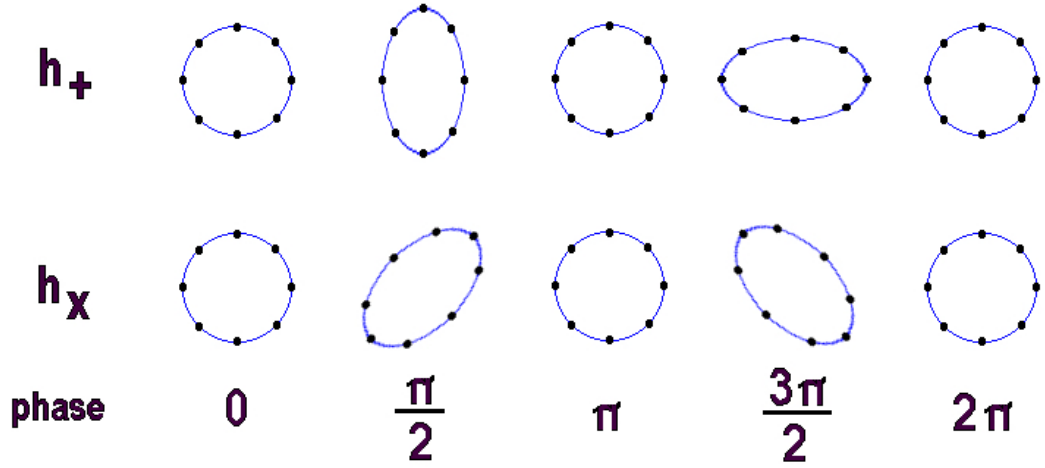


Figure 1.2: Illustration of the two gravitational wave polarisations. When a GW passes by, the space would be squeezed in one direction while stretched in the other direction, there are two polarisation modes, known as the plus mode denoted as  $h_+$ , and cross mode denoted as  $h_\times$ . Reproduced from [19]

## 1.2 Gravitational Wave from Compact Binary Coalescence

The GW signal emitted from compact binary coalescence (CBC) systems, which consist of two components of either Neutron Stars or Black Holes, are among the most promising signal sources. Binary neutron star systems have been observed, and astronomical phenomena like sGRB are believed to originate from such CBC events (see for example [20, 21, 15, 22, 23] and the references within).

In order to achieve detection in noisy data, one needs to compute the waveform of CBC events. For inspiral signals, the waveform is relatively easy to calculate, since the two compact objects are still relatively well separated. Thus it is still valid to approximate them as point masses, and the Post Newtonian (PN) approximation could be applied so that analytical expression could be achieved [17, 24]. The higher order PN approximation is used, the more time-consuming to generate the waveform, and more factors included [25]. However, in the merger stage when the compact objects get close enough, say, around the Innermost Stable Circular Orbit (ISCO) beyond which the system plunge in and quickly coalesce, then the PN can no longer track the waveform evolution, and more accurate simulation is required [26]. This could be achieved by adopting the results of Numerical Relativity (NR) [27]. It took researchers decades of effort to successfully simulate a binary black hole rotation without crashing the code with NR. However, in recent years, the huge advantage of NR has

been achieved, and we can expect to simulate the merger period accurately [26]. When the compact binary merges into one body, it will either form an unstable HyperMassive Neutron Star (HMNS), which will last for  $\sim 100\text{ms}$  before it collapses into a Black Hole (BH); or it will directly collapse into a BH, depending on the EoS of the Neutron Star (NS) [28, 14]. After the two compact objects have merged into one black hole, the asymmetric part will emit the ringdown signal, and eventually fade away [29, 30]. When the BH is formed, things are getting simpler again, since the Effective One Body (EOB) formalism could be applied to obtain the ringdown waveform [31, 32, 33, 26]. Still, such waveform generation would consume large amounts of computational time, so the current strategy is to generate a set of waveforms that covers the possible range of the parameter space. Later, the interpolation with such waveforms could be applied to obtain waveforms with a specific parameter set [34, 35, 36].

In the following section, we estimate the GW strength from a CBC inspiral with a simple heuristic order-of-magnitude estimation [3, 4]. For a binary with two components of mass  $m_1$  and  $m_2$ , the potential energy of the system is

$$U = - \int_{\infty}^R dr F(r) \sim - \frac{Gm_1m_2}{R}, \quad (1.6)$$

if we define the reduced mass  $\mu \equiv \frac{m_1m_2}{m_1+m_2}$  and the total mass  $M \equiv m_1 + m_2$ , then equation 1.6 can be rewritten as  $U \sim - \frac{G\mu M}{R}$ . On the other hand, a CBC would be stable due to the orbit circularisation as we will discuss in detail later. Applying the virial theorem, one can obtain

$$\langle T \rangle \sim \langle E_{kin}^{nonsp} \rangle \sim - \langle U \rangle \quad (1.7)$$

where  $T$  is the kinetic energy and the  $\langle \cdot \rangle$  represents average over time. Introducing  $\mathcal{M} = \mu^{3/5} M^{2/5}$ , and combining with the fact that the frequency of the system is  $f \sim \sqrt{M/R^3}$ , we can rewrite equation 1.3 into [2, 3, 17]

$$h \sim \frac{G^2}{c^4} \frac{\mathcal{M}^{5/3} f^{2/3}}{r}. \quad (1.8)$$

Assuming that the orbits are already circularised, we can approximate the orbit as circular. Further we can assume that the phase evolution is much faster than the amplitude evolution, or using the *stationary phase approximation*, namely the time evolution of the  $h(t)$  is simply

sin and cos functions to first order [2, 37, 38]. Then the luminosity of the GW will be

$$\begin{aligned}
 L_{GW}^{CBC} &\sim \frac{c^3}{32\pi G} 4\pi r^2 h^2 f^2 \\
 &\sim \frac{G^3 \mu^2 M^3}{c^8 R^5} \\
 &\sim \frac{G^3}{c^8} \mathcal{M}^{10/3} f^{10/3}
 \end{aligned} \tag{1.9}$$

We can observe that the GW luminosity is strongly dependent on the distance between the binaries, so the impact of gravitational radiation could be approximated by an impulse at the periapsis, which would shrink the semi-major axis while keeping the pericentre distance intact, effectively decreasing the orbit eccentricity. This explains the effect of orbit circularisation [39].

We can also deduce

$$\begin{aligned}
 \frac{dR}{dt} &= \frac{dR}{dE} \frac{dE}{dt} \\
 &\sim -\frac{R^2}{G\mu M} \frac{G^3 \mu^2 M^3}{c^8 R^5} \\
 &\sim -\frac{G^2 \mu M^2}{c^8 R^3},
 \end{aligned} \tag{1.10}$$

which points to the frequency evolution

$$\begin{aligned}
 \dot{f} &= -f \frac{\dot{R}}{R} \\
 &\sim \frac{G^2}{c^8} f \frac{M^2}{R^4} \mu \\
 &\sim \frac{G^2}{c^8} \mathcal{M}^{5/3} f^{11/3}.
 \end{aligned} \tag{1.11}$$

This implies that, for CBC systems, the evolution of frequency accelerates as the binaries approach each other more closely. Under such stationary phase approximation, the Fourier transform of the waveform  $\tilde{h}(f) \propto f^{-7/6}$  [38, 17, 2].

## 1.3 Detecting Gravitational Wave with Ground-Based Detectors

Currently the most sensitive observation of GW comes from ground based laser interferometer Gravitational Wave Detectors (GWDs), which essentially use laser interferometers to detect tiny strain of the space [3, 2]. As shown in figure 1.3, when a GW signal comes from

the zenith with proper polarisation, one arm will be shrunk and the other one will be stretched very slightly. The difference between the two armlengths will vary, and would contain the trace of GW signal, which can be detected using sophisticated innovations in technology. In the long wavelength limit where the wavelength is much longer than the detector length

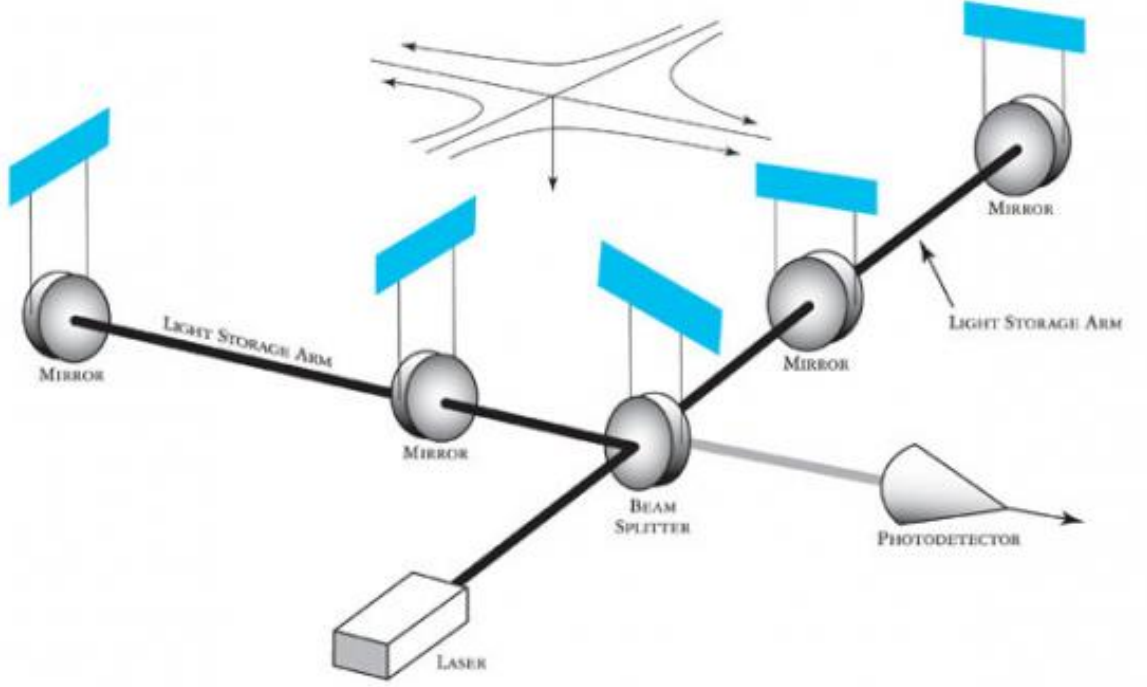


Figure 1.3: Illustration of the principle of laser interferometer gravitational wave observatory, the passing gravitational wave signal will induce difference between armlengths, and the laser interferometer will detect such tiny variation. Reproduced from [40]

scale, the coupling between the detector and the passing GW signal could be written as [4, 3]

$$\delta L(t) = \frac{1}{2} \mathbf{d} : \mathbf{h} \equiv \frac{1}{2} d^{ij} h_{ij}. \quad (1.12)$$

We apply the Einstein summation notation where repeated indices are summed over [1, 18]. The expression of the detector tensor  $\mathbf{d}$  differs for different configurations of detectors. For a laser interferometer with two arms perpendicular to each other, it can be written as

$$\mathbf{d} \equiv L(\mathbf{e}_x \otimes \mathbf{e}_x - \mathbf{e}_y \otimes \mathbf{e}_y).$$

We can express the strain amplitude from equation 1.2 as

$$\frac{\delta L(t)}{L} = F_+(\theta, \phi, \psi) h_+(t) + F_\times(\theta, \phi, \psi) h_\times(t) \quad (1.13)$$

where the *antenna pattern* of the detector  $F_+$  and  $F_\times$  for a source from sky location  $\theta, \phi$  and polarisation angle  $\psi$  [4, 3, 17]

$$F_+ \equiv \mathbf{d} : \mathbf{e}_+ = \frac{1}{2}(1 + \cos^2 \theta) \cos 2\phi \cos 2\psi - \cos \theta \sin 2\phi \sin 2\psi$$

and

$$F_\times \equiv \mathbf{d} : \mathbf{e}_\times = \frac{1}{2}(1 + \cos^2 \theta) \cos 2\phi \sin 2\psi + \cos \theta \sin 2\phi \cos 2\psi,$$

where  $:$  is the double dot product, meanwhile

$$h_+(t) \equiv \frac{2\mathcal{M}_z}{D_L} [\pi \mathcal{M}_z f(t)]^{2/3} (1 + \cos^2 \iota) \cos[\Phi(t)]$$

and

$$h_\times(t) \equiv \frac{4\mathcal{M}_z}{D_L} [\pi \mathcal{M}_z f(t)]^{2/3} \cos \iota \cos[\Phi(t)].$$

while

$$\Phi(t) = \phi_0 + 2\pi \int_{t_c} f(t') dt' = \phi_0 + 2\pi \int_{t_c} \frac{f}{df/dt} df$$

with

$$\begin{aligned} \frac{df}{dt} = & \frac{96}{5} \pi^{8/5} \mathcal{M}_z^{5/3} f^{11/3} \left[ 1 - \left( \frac{743}{336} + \frac{11}{4} \eta (\pi \mathcal{M}_z f)^{2/3} \right) + 4\pi (\pi \mathcal{M}_z f) \right. \\ & \left. + \left( \frac{34103}{18144} + \frac{13661}{2016} \eta + \frac{59}{18} \eta^2 \right) (\pi \mathcal{M}_z f)^{4/3} \right] \end{aligned}$$

according to [24, 37]. Here  $\mathcal{M}_z = \mathcal{M}(1+z)$  is the redshifted chirp mass,  $D_L$  is the luminosity distance,  $\iota$  is the inclination angle. We can thus conclude that the CBC signal is controlled by nine parameters, including the intrinsic parameters chirp mass  $\mathcal{M}$ , mass ratio  $\eta$ , inclination angle  $\iota$  which depends only on the source configuration, and extrinsic parameters  $\theta, \phi, D_L, \psi$ , an arbitrary reference time  $t_c$  and related reference phase  $\phi_0$ , which also relate to the observer [41, 2, 3]. The relationship between the three angles  $\theta, \phi$  and  $\psi$  is demonstrated in figure 1.4

## 1.4 A brief history of Gravitational Wave experiment

The first attempt to detect GWs was carried out in the 1960s. Weber built the first functioning gravitational wave detector (GWD), known as the “Weber bar”, which uses the whole bar to resonate with the passing GW signal [42, 43, 44]. As shown in figure 1.5, the passing GW will induce a slight deformation of the bar, with the change in the length  $\Delta l$  divided by the original length of the bar  $l$  defining the amplitude of the GW,  $h = \frac{\Delta l}{l}$ . If the frequency of the passing GW lies near the resonance frequency of the detector, the detector would be excited and resonance of the bar amplifies the signal. Similar configurations of GWDs were built

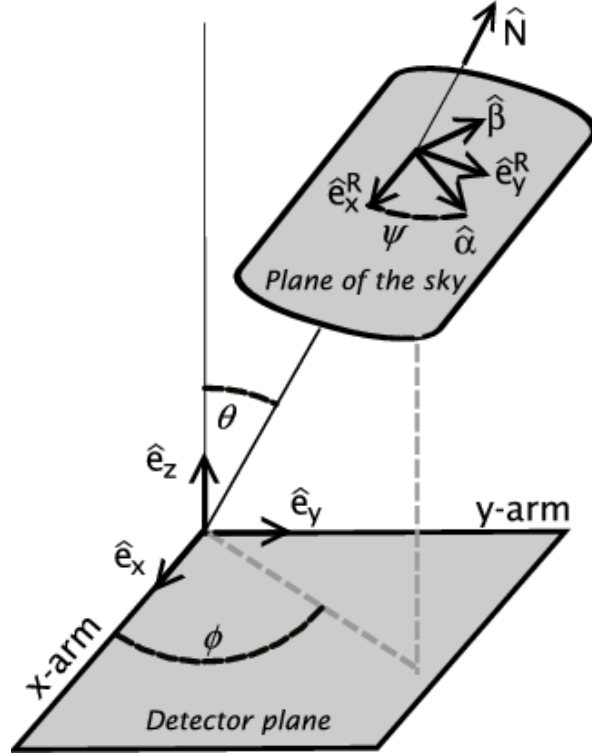


Figure 1.4: Illustration of the definition and relationship between the two directional angles  $\theta$ ,  $\phi$  and the polarisation angle  $\psi$ . Reproduced from [2]

throughout the world, including the Glasgow resonant detector [46].

The first direct detection was claimed by Weber [42, 43, 44]. However, following experiments failed to repeat such a detection [46, 45]. The current laser interferometers have reached a much higher sensitivity, while no single detection has been made yet [47]. Thus it's generally believed the events detected by Weber were more likely to be noise than astronomical events.

The first successful evidence of GW has been achieved by studying the binary Neutron Star discovered by Hulse and Taylor [7, 5, 6], who were awarded the 1993 nobel prize “for the discovery of a new type of pulsar, a discovery that has opened up new possibilities for the study of gravitation”. [48] As shown in figure 1.1, the measured orbital evolution of this system match extraordinarily well with the theoretical prediction of gravitational radiation [5][6],

$$\dot{P}_{b,GR} = -\frac{192\pi G^{5/3}}{5c^5} \left(\frac{P_b}{2\pi}\right)^{-5/3} (1-e^2)^{-7/2} \times \left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right) m_p m_c (m_p + m_c)^{-1/3} \quad (1.14)$$

where  $\dot{P}_{b,GR}$  is the theoretical decay rate for the orbit period according to GR,  $P_b$  is the orbit period,  $e$  is the eccentricity,  $m_p$  and  $m_c$  are the mass of the pulsar and the component separately and could be accurately determined via independent observation. The observation





Figure 1.5: Bar detector in Munich together with H. Billing. Reproduced from [45]

of such BNS thus assures the existence of GW, and huge interest was triggered to search for the GW signal directly.

Bar detectors are naturally limited to detect GW signals with  $h > 10^{-21}$  [2], while the existing interferometers have reached sensitivity higher than such a limit. As shown in figure 1.6, for bar detectors like Auriga the sensitive frequency region is very narrow, and can only reach a sensitivity around  $10^{-21}$  [49]. So in recent decades, the interests of detecting

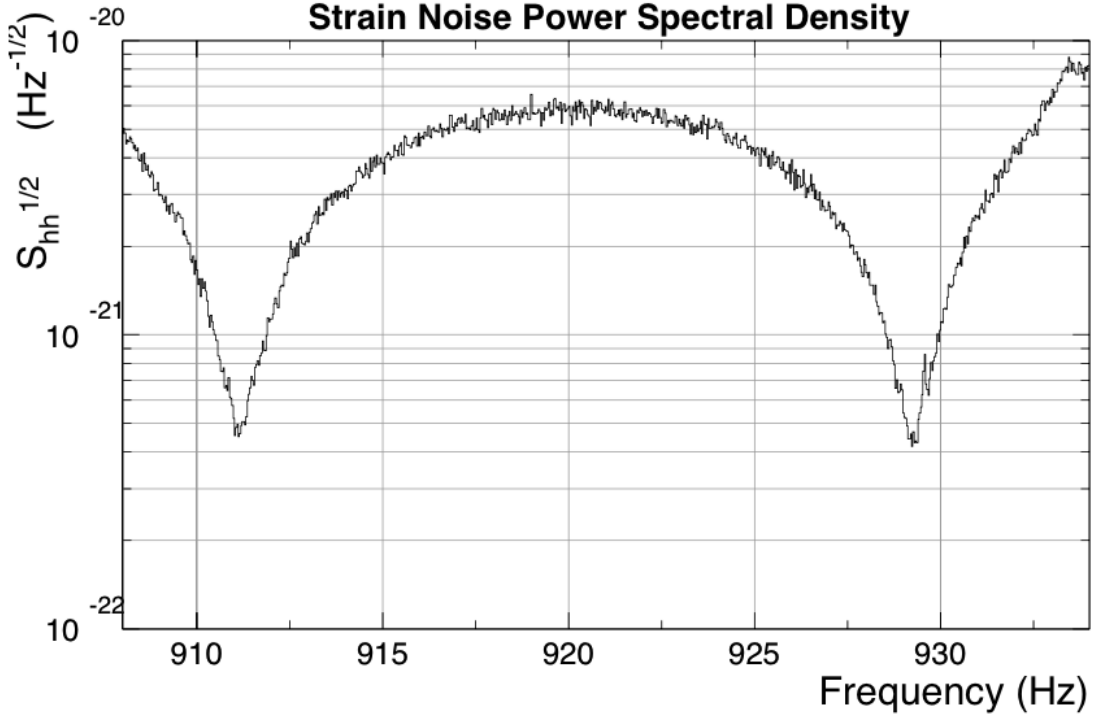


Figure 1.6: The power spectral density for the bar detector Auriga. One can notice that the most sensitive part can only reach around  $10^{-21}$ , and the sensitive frequency band is very narrow. Reproduced from [50]

gravitational waves have shifted from bar detectors towards laser interferometers. In 1992, the building of the Laser Interferometer Gravitational-Wave Observatory (LIGO) was started, and the initial generation of LIGO started operation in 2002 [51]. A European counterpart with a similar design of laser interferometer, known as Virgo, started operating in 2007, and this global network of LIGO-Virgo has existed since then [52].

No confident detections of any kind of astronomical signal were detected during the operation of these initial generation detectors so far [47]. However, even the non-detection could lead to scientific conclusions. For example, no GW signals were discovered for the Vela pulsar [53] nor Crab pulsar [54], which puts a solid upper limit of the GW strength of such young pulsars. This itself already put some constraints on the NS ellipticity [55].

Since the end of observation by initial LIGO, the upgrading of this initial generation into second generation GWDs, known as Advanced LIGO (aLIGO) [56] and Advanced Virgo (AdV)

[57] respectively, has been under construction. These next generation GWDs are expected to gain an increase of sensitivity by about 10 fold in all frequencies. Since the detection volume is proportional to the third power of the detectable distance, such an improvement translates into an increase of detection efficiency of approximately three orders of magnitude [58], as illustrated in figure 1.7, A new set of advanced GWD will be built in India, known

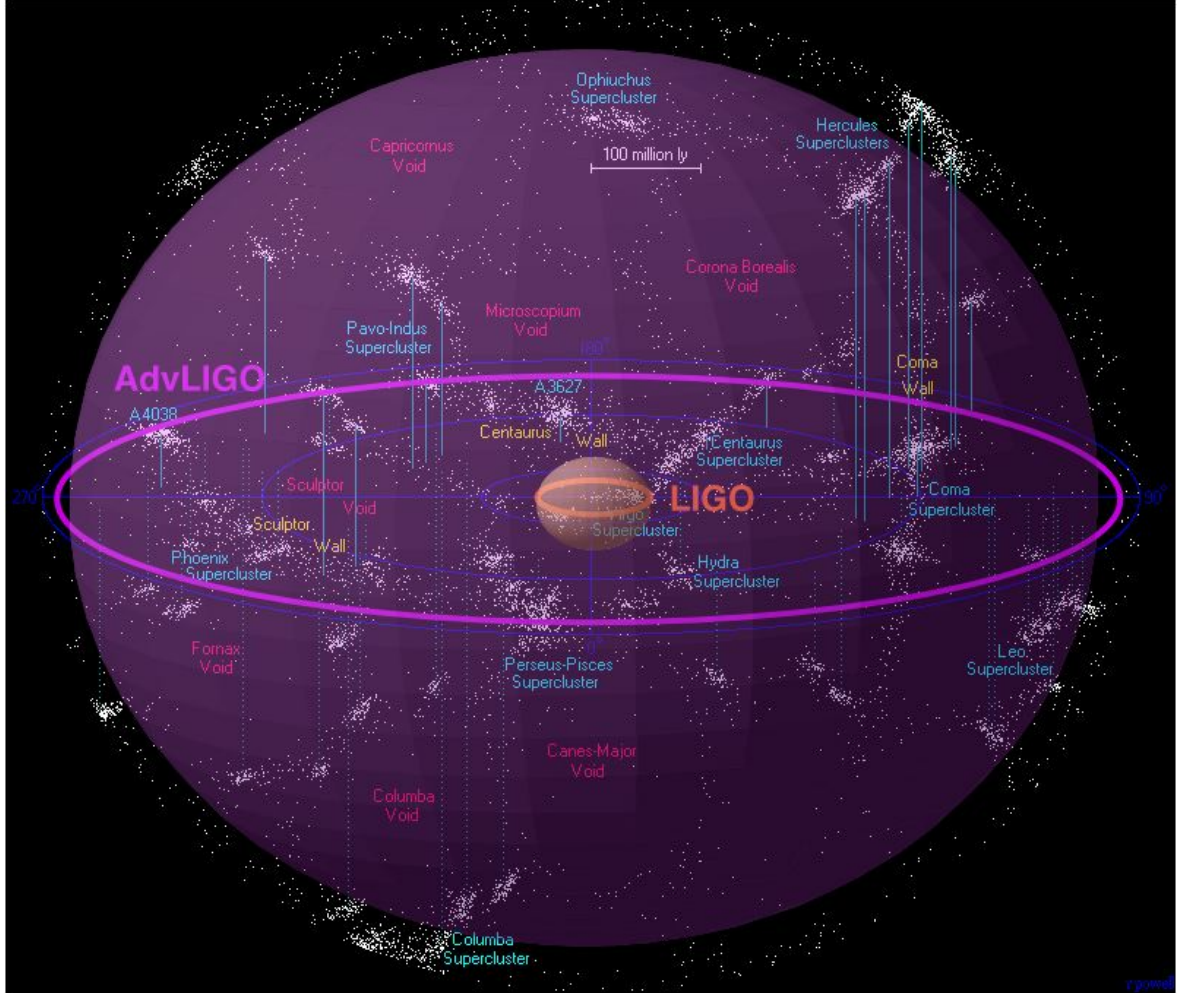


Figure 1.7: Sensitive volume for initial LIGO (central sphere) and advanced LIGO (outer sphere). A factor of 10 improvement of sensitivity in  $h$  will lead to an increase of a factor of 10 in distance. As sensitive volume would be proportional to the cube of the sensitive distance, advanced LIGO is expected to enhance the detection probability by  $\sim 1,000$  from the initial LIGO level. Reproduced from [59]

as the Indian Initiative in Gravitational-wave Observations (IndIGO) [60]. Huge progresses has been made towards the full upgrade of LIGO detectors, which is expected to be finished in near future.

In the mean time, the Japanese Large-scale Cryogenic Gravitational Wave Telescope Project, or KAGRA, is under construction [61]. Combined with the UK-Germany prototype detector GEO, a global detector network would provide a robust hardware facility to detect GW [62].

There are various noise sources for a GWD, among them the most outstanding ones are gravity gradient noise at very low frequency, quantum effects of light pressure at low frequency, thermal noise in the middle, and shot noise at high frequency [2, 3]. Figure 1.8 demonstrates the anticipated noise budget for aLIGO, indicating the frequency dependence of these different effects.

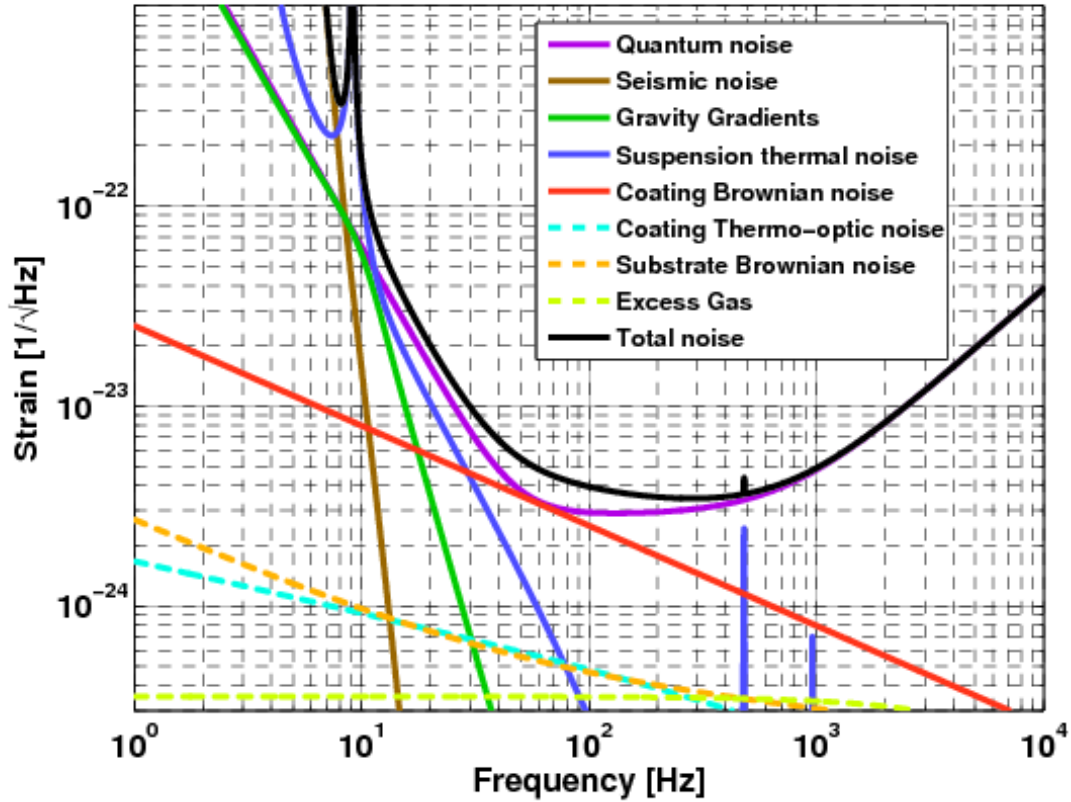


Figure 1.8: Expected noise budget of the advanced LIGO, with different components labeled individually. In the low frequency range ( $f < 10\text{Hz}$ ), seismic noise, gravity gradient noise and suspension thermal noise dominate, in the higher frequency the quantum noise would contribute mostly to the noise budget. Figure reproduced from [63]

The future generation of GWD is already in discussion, and the proposed Einstein Telescope (ET) detector, which has an underground equilateral triangle configuration, will have further increase of sensitivity by about another factor of ten [58, 64]. Meanwhile, the extension of the undergoing aLIGO to the  $A+$  (or Quantum LIGO) has also been under serious consideration as a near term upgrade [65].

Meanwhile, longer armlengths will shift the ‘sensitivity bucket’, namely the most sensitive range of the detector, towards lower frequency, which corresponds to  $\sim 100 - 1000\text{Hz}$  in figure 1.8. Plans were also proposed to build detector in space. The proposed Laser Interferometer Space Antenna (LISA) project aimed to launch three satellites into space to form an equilateral triangle interferometer. The proposed armlength of LISA will be 5 million

kilometres long, this enables the accurate observation of GW signal within  $0.03 - 0.1\text{Hz}$  range [66].

Unfortunately, the funding for research of LISA was cut down, thus there is no definite launch plan for it in near future. The current proposal is to build a down-scaled space-borne GWD, Evolved Laser Interferometer Space Antenna (eLISA) [67]. However, the LISA pathfinder, which is a single satellite aimed to test the technology applied on LISA, has already been integrated and is scheduled to be launched in 2015 [68]. If it succeeds in demonstrating the safeness and reliability of the technology, there's still hope to continue and extend the study of eLISA.

Other proposed future projects include the DECI-Hertz Interferometer Gravitational wave Observatory (DECIGO) [69], which focuses on the frequency range between LISA and aLIGO, and Big Bang Observer (BBO) [70], which is proposed to be the successor of LISA, equipped with enough sensitivity to observe the GW emitted from shortly after the big bang.

Other methods of GW detection includes the Pulsar Timing Array (PTA), which takes advantage of the accurate timing of pulsars to capture small fluctuations in the Time Of Arrival (TOA) of steady pulsars; and B-mode polarisation of CMB, which could trace back to the end of inflation [71]. As noted earlier, the BICEP2 team originally claimed that they detected such B-mode polarisation in CMB signal; however, this signal turns out to be very likely coming from Galactic dust, which was not adequately considered by the BICEP2 team [10]. GWs with frequency  $100\text{MHz}$  to  $100\text{GHz}$  and even higher could be detected with novelly designed experiment, the potential source being relic GWs [72]. Interesting upper limits on GW strength could also be given by monitoring the timing of the Global Positioning System (GPS) system [73]. Progress has been made to use the whole Earth as a resonant detector, and upper limits on GW signals could be achieved by checking seismic monitoring data [74].

## 1.5 Astronomical Sources for Ground-based Gravitational Wave Detectors

In addition to the tremendous efforts over recent decades to build ever more sensitive detectors, the equally large amount of effort on the data analysis side is also vital in order to create a robust analysis frame in which potential detections can be identified, and also to make sure that we can properly understand the physics that such detections can reveal [41, 75, 76, 77, 78].

The first question that theoretical researcher should answer is how likely is it that we could detect a GW event, provided we could reach a certain sensitivity. This involves understanding

the possible mechanisms that could emit large amounts of GW radiation [4, 2].

Currently, we believe that the CBC process of binaries comprising binary NS, binary BH, or a mixture of both, would be the most promising physical mechanism to generate large amounts of GW radiation in the most sensitive frequency range of the ground based GWD [79]. The binary system will rotate around the centre of mass, while the GW radiation will carry the angular momentum and kinetic energy away, shrinking the binary orbit closer. The binary will rotate faster and the GW radiation will become stronger, until they reach the ISCO, where the binary will plunge in and eventually merge into a BH, causing a ‘chirp’ signal. The peak frequency is directly related to the ‘chirp mass’, roughly speaking, the heavier such binary system is, the lower the chirp frequency would be [2, 3]. BNS systems are expected to be the best candidate, while binary Black Hole (BBH) systems are generally heavier, thus only the merger and ringdown part of the waveform would be resolved for these systems [80].

CBC signals, especially BNSs, are expected to be responsible for triggering the sGRBs, thus we can have rough estimation on the astronomical event rate of such events [81, 82]. However, the beaming angles of such events are highly uncertain [83, 84, 85], which causes the large variance in the estimation [86]. The other independent method to estimate such event rate would be using population synthesis methods to predict the expected fraction of compact binaries that could lead to a CBC. Since there are large uncertainties in both methods, the estimation of the rate is very vague so far, and the “optimistic” and “pessimistic” estimations differ by three orders of magnitude. In terms of the advanced LIGO-Virgo network, the “realistic” estimation of the rate of BNS signals would be around 40 events per year [79]. Studies suggest that apart from sGRBs, kilonovae are also excellent electromagnetic (EM) counterparts of the BNS events [20, 87]. Once a BNS event is detected together with an EM signature, we can assure the detection is indisputable, and also the information combined from both GW and EM channel would make the GW cosmology possible [88, 15, 22, 89].

The Intermediate-Mass Black Hole (IMBH) has been proposed to explain the growth of supermassive black hole and is also predicted by the correlation of the BH mass and the stellar environment. The GW observation of an Intermediate-Mass Ratio Inspiral (IMRI), or the inspiral of a binary system with one IMBH and a stellar mass NS or BH, and the merger of two IMBHs can lead to deeper understanding of the growth of supermassive Black Hole (SMBH) and the interaction of the central black hole and the surrounding stars’ activity, as well as putting solid constraints on the gravitational physics [90, 91, 92].

The existence of BNS is confirmed from both observation [7, 5, 93] (and reference therein) and population synthesis [94, 95]. There’s still lacking convincing evidence for the existence of Neutron Star-Black Hole (NSBH) or BBH systems, mostly because lacking of traditional observing channels. For IMBH, there is evidence favouring their existence, but the evidence

IFO	Source	$\dot{N}_{low}$ yr <sup>-1</sup>	$\dot{N}_{re}$ yr <sup>-1</sup>	$\dot{N}_{high}$ yr <sup>-1</sup>	$\dot{N}_{max}$ yr <sup>-1</sup>
Initial	NS-NS	$2 \times 10^{-4}$	0.02	0.2	0.6
	NS-BH	$7 \times 10^{-5}$	0.004	0.1	
	BH-BH	$2 \times 10^{-4}$	0.007	0.5	
	IMRI into IMBH			< 0.001	0.01
	IMBH-IMBH			$10^{-4}$	0.001
Advanced	NS-NS	0.4	40	400	1000
	NS-BH	0.2	10	300	
	BH-BH	0.4	20	1000	
	IMRI into IMBH			10	300
	IMBH-IMBH			0.1	1

Table 1.1: Estimated detection rates for the initial LIGO and the advanced LIGO detector era for different sources. Notice the huge uncertainty in the estimation, with the most realistic estimation for the binary neutron star merger event rate estimated around 40 per year. Table adapted from [79].

is not decisive. In table 1.1, the detection rates for these different types of events are shown. The most optimistic and most pessimistic estimation for the rate differs by several orders of magnitude. We can also notice that the non-detection of CBC events during the initial LIGO can be well explained by the theoretical prediction, and there's promising event rate for the advanced detector era [79].

CBC events are expected to be rare, but relatively bright in GW emission. However, another potential detectable source for GW radiation would be the continuous wave (CW), which is more quiet and more nearby [96]. For NSs with a mountain on the surface, there would be an quadrupole radiation from the bump [2]. Although the GW strength of such source is expected to be much lower than the CBC signal, since there are large amount of NSs inside the Galaxy, such sources are expected to be much closer to the Earth [3]. The CW signals are also expected to last for a long time, making it possible to accumulate very long signal and thus accumulate the Signal-to-Noise Ratio (SNR) [97]. The strength of the GW is related to the size of the bump on the surface of NS. If we ever detect any loud signal from CW sources, it would imply that the EoS of NS must be stiff enough to support a high enough mountain [55]. The 5<sup>th</sup> and 6<sup>th</sup> scientific data S5 and S6, being the most sensitive data recorded by initial LIGO, showed no conclusive evidence of the existence of CW signal, putting a solid upper limit on the strength of the corresponding GWs [53, 54]. Notice that both magnetic brake and CW radiation could lead to the observed long term spin-down of pulsars, while young pulsars show very rapid spin-down. Prior to the non-detection of CW from initial LIGO, no solid evidence could reveal the corresponding mechanism for such observed spin-down. And now the conclusion can be made that at least for pulsars like Vela and Crab pulsar, the spin-down corresponding to CW contributes no larger than 45% [53].

One could also detect the stochastic noise by correlating multiple detectors [98]. When there is more than one source in a given frequency bin, one can't distinguish them from noise, and this would happen if the detector is in low enough frequency and/or sensitive enough, like we expected in eLISA and ET [2, 99]. This source of signal behave just like noise in one detector, changing the noise behaviour, and Parameter Estimation (PE) can not be accurately performed for each individual source [100]. However, this source is astronomical and thus could distinguished from local background noise. With data from multiple detectors, one can possibly construct the so-called null stream, which by design contains no signal at all. This is effectively “turning off” all astronomical sources. And by comparing the Power Spectrum Density (PSD) of the null stream with individual detectors' PSD, the contribution of a so-called stochastic background could emerge and one can study the property as a population. For eLISA, the population of Galactic binary White Dwarf (WD) and inspiral of binary SMBHs would become a stochastic source, while for ET, thanks to its sensitivity, the remote inspiral signals are expected to become stochastic noise [99].

The remaining sources are classified as burst signals, which refers to all short duration signals that are either not well-modelled or the corresponding mechanism is badly understood [101, 2].

Bad modelling leads to an inefficient usage of the data, thus the burst signal must be loud enough to be detectable. This requires the source to be either very close to the GWD, or being intrinsically loud, or a combination of both.

One such known candidate is core-collapse supernova (CC-SN). In order to produce a CC-SN, the mass distribution can not be isotropic, thus making it a promising GW source. However, the modelling of the process of CC-SN is a puzzling problem, and we are far from fully understanding it [102, 12]. The other issue relating to CC-SN is that it's intrinsically dimmer than a CBC signal in the GW channel [102]. So we only expect to detect very local signals. This leads to us a benefit that the detection of the EM counterpart is almost certain [75, 103].

The other potential burst sources includes r-mode instability of NS [104], Type Ia supernova [12], cosmic string cusp [3] and other unknown mechanisms. Just as the operation of radio telescopes led to the discovery of pulsars and quasars, we can't predict what new phenomena the GW window can bring us. For these unknown sources, it's simply impossible to predict the waveform, and we can only deal them as burst signals where no model is used as an auxiliary.



## 1.6 The Role of Data Analysis towards Detecting Gravitational Waves

The lowest order of gravitational emission is generated from a quadrupole formalism [24]. Compared with dipole radiation like EM radiation, this is much weaker. The gravitational constant  $G$  itself is also many orders of magnitude smaller than the fine structure constant  $\alpha$ , making the GWs only very weakly couple with matter, which further explains the extreme difficulty of their detection [3, 4].

The current laser interferometers are already very sensitive so that even clouds passing by or wind blowing could induce a detectable signature as the gravity gradient noise [105]. However, they are still not yet sensitive enough make a direct detection [53, 47, 106]. Even with the advanced detectors like aLIGO and AdV, the first detections are expected to have relatively low SNR [75]. How to pick signals up from the noisy background is the challenge left to be tackled for the data analyst [77, 78, 107, 108].

For well modelled signals like inspiralling CBC and CW, the GW waveform can be determined by the corresponding parameters like the chirp mass of binary system, or the frequency for a pulsar. By coherently integrating the data over a long enough time, the signal can stand out from the noise, thus increasing the SNR.

As illustrated in chapter 1.1, the CBC signal could be decomposed into three distinct regions, *i.e.* the inspiral, merger and ringdown [2, 4]. The inspiral stage is the longest, and any misalignment between the template and the actual data would lead to cancellation, so it puts best constraints on the phase parameters like chirp mass. The merger stage, although short in time, is the most violent stage, and thus contributes significantly to the SNR [4, 26]. The ringdown stage is also important as the perturbation of the geometry of spacetime around the BH would be an ideal test field for gravity theories like GR [31, 32]. However, since they are generated by three different methods, these stages need to be carefully spliced into one waveform, to avoid a potential shift of phase for the waveform [26]. As for physicists in general, what's most interesting is what physics we can learn from analysing the GW data. So given the detected data, one wants to know what the corresponding physical system should be, and how likely such a parameter combination is. This question relies upon the application of statistics to the data, especially Bayesian Inference (*c.f.* chapter 2) [47].

The important role that data analysis plays here is two-fold, namely the detection and measurement [16]. Both aims could be fulfilled by performing Bayesian Inference, but the aim of detection places a much stronger constraint on the speed of data processing, thus limiting us to applying Frequentist methods (*c.f.* chapter 6).

The current pipelines first generate a sparse template of waveforms, and then cross-correlate on the data with each waveform [108, 109]. Astronomical CBC signals will trigger a large

enough response of the data, such triggers with SNR larger than a certain threshold would be recorded as potential candidates for astronomical signals. However, since the data was matched with a large number of templates, thus even though the SNR threshold is supposed to shield a large fraction of noise-mocked signal, there are still a large number of false alarms. The current strategy of picking up the more plausible candidates is by looking at time and mass coincidence between multiple detectors [108, 109]. According to GR, GW signals are supposed to travel at the speed of light, and the different detectors should record the time of the CBC event trigger time differing by the time of flight that the GW takes to travel across multiple detector [106]. The underlining philosophy is that the noise among different detectors should be independent, so it is very rare for random noise to coincide in time and in parameter space, while the astronomical signal should pass such threshold easily [78, 107, 77]. However, at this stage, some amount of noise could still pass the threshold and mimic the signal. So there is a final threshold on the combined SNR: coincidences that pass such a threshold would be regarded as significant enough to make a confident announcement about the detection. However, the understanding of the real meaning of significance is still not solid enough, motivating us to learn more about the significance of the GW coincidence events (*c.f.* chapter 6).

Once a coincidence is significant enough, the corresponding data will be further passed to the PE pipelines for detailed analysis [17, 76, 47, 41]. The current bottleneck is the expensive computational consumption during such a process: the full PE of the inspiral signal could take as long as one week [17, 110]. The process of PE calculates the source parameters (or their posterior distribution to be precise) for some selection of points. Efforts have been made to both decrease the number of points to be computed [41, 111], and the time spent to calculate the statistic for each point [110, 35]. However, we should note that for advanced detectors, when the low frequency sensitivity largely increased, the low frequency inspiral signal could be observed for much longer time, thus making the process of PE even more challenging [110].

## 1.7 Physics and Astronomy with Gravitational Waves

What can GW provide to physics and astronomy?

First of all, our understanding of gravity is still not mature, as currently the most successful physics in large scale and heavy mass, GR, and the most successful physics in small scale, Quantum Mechanics (QM), are not consistent with each other [11]. And the gravity as a fundamental interaction, can hardly be unified with the remaining three interactions. The precise value of the gravitational constant is even under debate for the fourth significant figure, as shown in figure 1.9 [112].

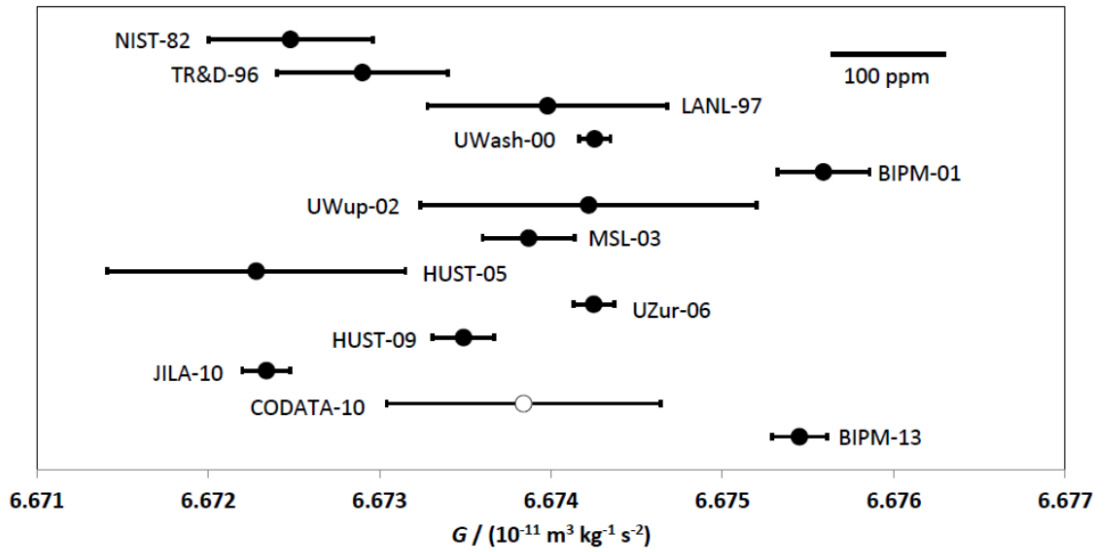


Figure 1.9: Measured value of the gravitational constant  $G$  from different research groups. Notice the values differs at the fourth significant figure. Picture modified based on lecture from Jim Faller

Meanwhile, GR deviates from Newtonian gravity in strong fields and at high speed [1, 18]. Up until now, no direct measurement for GR effect had been made in a strong field environment [113]. The strength of field of WD is only  $\sim 10^{-4}$ , for NS it's about  $\sim 0.1$  [2]. For EM astronomy, the detection strongly relies on photon radiation, which either happens in the outer region, or the environment has too thick optical depth that the photon from inner region would lose the information related to the strong field [14, 103]. GW radiation, on the other hand, only weakly couple with normal matter, so the 'optical depth' of the environment is very thin, allowing the GW signal easily to carry the inner information towards the detector [3]. This optical depth argument could also be applied to the big bang. Prior to the emission of CMB, the optical depth is too thick for EM signal to travel [114]. However, GW signal can be traced back until the end of inflation, leading to the potential direct study of the big bang and inflation [115, 115, 116].

With the detection of GW, we can also solve various astronomical puzzles. The actual mechanism of galaxy formation is under heated debate, while SMBH certainly play an important role, but the detailed process, especially when two galaxies merge, and the question of how the two SMBHs lose their angular momentum and merge into one single SMBH, is still unknown [117]. The Gamma Ray Burst (GRB) phenomenon, especially the sGRB, is also not fully understood [118, 119]. Although some knowledge about them has been gathered, and the link between GRBs and corresponding celestial objects were established, the central engine of the GRB still remains unclear [86, 120, 37, 21, 87]. The binary evolution of stars, and especially the end of such evolution, also remains a puzzle, causing the huge uncertainty in the estimation of CBC event rates [95, 79]. With a large enough number of GW detections,

especially CBC signals, the CBC event rate could be determined accurately, thus shedding some light on the binary evolution problem [121, 117, 39]. Also, a more detailed study of the waveform, especially the merger and ringdown waveform could distinguish the EoS of NSs [32, 31].

The combination of multiple channels, including EM, GWs and neutrinos, opens a field of multi-messenger astronomy [14, 122, 123, 124]. If a nearby massive star undergoes a supernova (SN) explosion, we can study the physical process of the outer region by collecting the EM signal, while the GW and neutrino can carry detailed information about the inner region, enabling a much more comprehensive understanding of the SNe [12]. However, it's very likely that the EM counterpart of a CBC event is weak, so that the EM telescopes need to know where to point to in order to find them [75]. This requires the data analysis of GWDs to perform rapid sky localisation. Normally this is achieved by using the difference in TOA for the data, assuming the signal to have a flat wavefront and that the difference is caused by the fact that wavefront reaches different GWD sites at different times. Taking the factor that GWs travel at the speed of light, one can reconstruct the sky location of the source with trigonometry. Such strategy relies on a widely separated GWD network, with more than two detectors in operation [75, 125].

The other significant potential application of GW and multi-messenger astronomy is to study cosmology with GW [15, 126, 127]. One of the usual methods to study the cosmology is to measure the distance and redshift of a group of standard candles, which leads to the accurate description of the redshift evolution history, and thus to decipher the cosmology [128, 129]. However, this process may require closer objects to calibrate the farther object, and step by step towards very distant cosmological objects. In this process, the calibration error will propagate all the way down to the distance of standard candle [130].

On the other hand, CBC signals, containing both strength information and the chirp mass information by themselves, represent a self-calibrated source, so one can determine the distance of a CBC source without the error propagation of multiple levels of calibration [15]. If the EM counterpart of such events were discovered in the mean time, one can accurately measure the redshift of such event, thus fulfilling all essential elements to perform GW cosmology [15, 131, 132].

To achieve that, the ability of detecting much more and much further events is needed, echoing the request to build future generation GWDs like ET [58, 64, 99]. For such detectors, thanks to the excellent performance of low frequency, a CBC signal could stay in band for a very long time, and the huge amount of signals makes it even possible that for every moment, the data contains detectable signals [99]. With such a rich collection of sources, one can perform a more robust statistical study, and a large number of phenomena could be studied, like gravitational lensing and star formation history across different periods

[116, 133, 134, 95, 135].

These scientific objectives all rely on a careful choice of the sites for the GWDs [136, 137]. Current detectors are actually all located in sub-optimal regions, as they are too close to the seashore, and the micro-seismic noise could constraint the limit of noise background (*c.f.* chapter 5). The future generation GWD are still in the early design stage and the site selection is under serious consideration. In this process, one should always bear in mind that since GWD has an all-sky response, so a global GWD network is more than a simple summation of multiple components [136]. Also, one should consider the factor that the real world is full of uncertain factors, so the detectors' site should be as flexible as possible.

## Chapter 2

# Applying Bayesian Inference to Gravitational Wave Data

As we described before, subject to numerous limits, we expect GW signals to be quite weak (*e.g.* [4, 3, 2, 97, 98]). With the first detected signal expected to have an SNR of merely around 10, data analysis techniques are required to obtain useful physical knowledge from the weak raw data [3, 2]. Among the various potential GW sources, CBCs, especially BNS inspiral signals are expected to be best understood [26, 109, 138]. Given a set of physical parameters to describe the CBC systems, the phase and amplitude evolution of the GW waveform in the inspiral stage can be predicted with high precision, under the PN scheme (see for example [139, 24, 140, 141, 142] and the references within). With the knowledge of the noise, one can calculate the probability to obtain such an observed data stream from a certain parameter set [41, 76, 143].

For a CBC system, especially for the inspiral stage where the physics is best understood, the evolution of the system can be efficiently computed with high accuracy, thanks to the formalism of PN [17, 144, 25, 145, 146]. By filtering the data with such predicted waveforms, a high response would be initiated if a signal of this form is embedded within the data. This method is known as matched filtering [29, 147, 148, 138].

The matched filtering strategy is applied to the problem of detecting and estimating the parameters of a CBC signal, while Bayesian Inference also plays an important role in the parameter estimation of CBC signals [99, 149, 47, 150, 16]. In other words, matched filtering is applied within a Bayesian context. We discuss these methods more thoroughly in the following sections.

## 2.1 Matched Filtering

In order to make the detection of gravitational waves, matched filtering, or cross-correlating the expected waveform to the observed data, is adopted. The basic logic of matched filtering is to check whether the data is consistent with an astrophysical waveform embedded in the noise with expected characteristics.

For a stationary random process  $n(t)$ , the *Power Spectral Density*, or PSD, describes the variance of the random process at different frequencies. The truncated Fourier Transform of a random process is defined as

$$\tilde{n}_T(\omega) = \frac{1}{\sqrt{T}} \int_0^T dt e^{-i\omega t} n(t), \quad (2.1)$$

and the long term average of the square of truncated Fourier Transform

$$S_h(\omega) = \lim_{T \rightarrow \infty} \mathbf{E} [|\tilde{n}(\omega)|^2] \quad (2.2)$$

is defined as the PSD [17, 151, 16].

Furthermore, we can define the autocorrelation function  $C_n(\tau) = \mathbf{E}[x(t)x^*(t-\tau)]$ . According to Wiener-Khinchin theorem, the one-sided PSD is just the Fourier transform of  $C_n(\tau)$

$$S_h(f) = \int_{-\infty}^{\infty} C_n(\tau) e^{2\pi i f \tau} d\tau, \quad (2.3)$$

where  $f > 0$  [2, 3, 16, 151].

For the recorded data  $d(t)$ , we expected it contain both background noise  $n(t)$  and the foreground gravitational wave signal  $h(t)$ ,

$$d(t) = h(t) + n(t)$$

although for most cases the signal is much too weak to be detectable, the data could be treated as noise.

If we have discrete samples at  $t_i = i \times \delta t, i \in \mathbb{N}$ , then it's straightforward to show that the probability of getting data  $d(t)$  sampling from the random process represented by  $S_h(f)$  is [16]

$$P(d|S_h(f)) = \frac{\exp(-\frac{1}{2} \mathbf{d}^T \mathbf{C}^{-1} \mathbf{d})}{[2\pi \det(\mathbf{C}_n)]^{1/2}}, \quad (2.4)$$

where  $C_{n,ij} \equiv C_n[(i-j)\delta t]$ , and  $\mathbf{C}^{-1}$  is just the inverse matrix of  $\mathbf{C}_n$ . In the limit  $\delta t \rightarrow 0$ ,

observing time  $T \rightarrow \infty$ , we can express [16]

$$\begin{aligned} \mathbf{d}^T \mathbf{C}^{-1} \mathbf{d} &= \int_{-\infty}^{\infty} df \frac{\tilde{d}^*(f)\tilde{d}(f) + \tilde{d}(f)\tilde{d}^*(f)}{S_h(f)} \\ &= 4\Re \int_0^{\infty} df \frac{\tilde{d}^*(f)\tilde{d}(f)}{S_h(f)}. \end{aligned} \quad (2.5)$$

Where  $\tilde{d}(f)$  represents the Fourier Transform of  $d(t)$  [16]. We can thus define the scalar product of  $(g|h)$  as [2, 17, 3, 16]

$$\begin{aligned} (g|h) &= 2 \int_0^{\infty} df \frac{\tilde{g}^*(f)\tilde{h}(f) + \tilde{g}(f)\tilde{h}^*(f)}{S_h(f)} \\ &= 4\Re \int_0^{\infty} df \frac{\tilde{g}^*(f)\tilde{h}(f)}{S_h(f)}. \end{aligned} \quad (2.6)$$

We can further construct the SNR of a signal as

$$\text{SNR} = \frac{(h|d)}{\sqrt{(h|h)}}$$

where  $h$  is the GW signal, and  $n$  is the noise [2, 17, 3, 16]. Since expected value for  $(h|n)$  is zero, so the expected value of the SNR can be re-expressed as  $\sqrt{(h|h)}$ . By applying the Cauchy-Schwarz inequality, one can prove that when the actual signal is  $h$ , the SNR value will be maximised and it constructs an optimal filter such that  $\rho_{\text{opt}} = \sqrt{(h|h)}$  is then defined as the *optimal SNR* [152, 38, 17].

On the other hand, if we want to perform PE on the GW data, we need to first construct the likelihood function  $P(d|\boldsymbol{\mu})$ , namely, the probability of obtaining data  $d$ , given the fact that the GW signal is determined by a set of parameters  $\boldsymbol{\mu}$  [41]. This is equivalent to the probability that the residual after removing the expected signal with  $\boldsymbol{\mu}$  from observed data, or [16, 3, 17]

$$P(d - h(\boldsymbol{\mu})|\text{Noise}) = \exp\left[-\frac{1}{2}(d - h|d - h)\right]. \quad (2.7)$$

With the linearity of the Fourier Transform, one can expand equation 2.7

$$\begin{aligned} P(d - h(\boldsymbol{\mu})|\text{Noise}) &= \exp\left\{-\frac{1}{2}[(d|d) + (h|h) - 2(h|d)]\right\} \\ &\propto \exp\left\{-\frac{1}{2}[(h|h) - 2(h|d)]\right\} \end{aligned} \quad (2.8)$$

The maximum of likelihood happens when the partial derivative of the likelihood over any parameter is zero, or [16]

$$\frac{\partial \log(\mathcal{L})}{\partial \mu} = 0. \quad (2.9)$$



It's not hard to rewrite this condition as [16]

$$\left(\frac{\partial h}{\partial \mu} \middle| d - h\right) = 0. \quad (2.10)$$

## 2.2 Bayesian Inference

In the field of statistics, there are two main branches known as frequentist and Bayesian [153, 154, 155]. They differ fundamentally in the definition of probability. For example, it is meaningless to discuss the probability for a single event in frequentist language, without introducing a hypothetical identical independent distribution [156]. Frequentists can of course predict the probability of obtaining one value before throwing the dice, but the correct expression would be “if the dice would be thrown for infinite times under identical and independent circumstances, what's the probability of getting one, or, what's the fraction of circumstances getting one out of all experiments”[156, 155]. Frequentist analysis agrees with common sense, but the meaning is not straightforward, and needs appropriate interpretation [156, 157].

Bayesian Inference, on the other hand, gives a straightforward definition of the probability, and it is meaningful to ask what is the probability of a single event in a Bayesian context (readers are encouraged to read this very interesting blog [158]). The benefits from adopting a Bayesian Inference approach also apply to the fact that the previous knowledge could be taken into consideration as the prior. In cases where the background knowledge is well understood while the data are of bad quality, then the final estimation, or the posterior, will be dominated by the prior. In cases where data are of excellent quality, the prior would be updated by this observation. So today's posterior could be tomorrow's prior [156, 157].

There are cases where one is more interested in certain subset of parameters than others, such uninteresting parameters are called *nuisance parameters*, like phase in the CBC signal. Bayesian Inference can also provide a convenient method to marginalise over nuisance parameters, which produce conclusions independent of any specific choice of nuisance parameter. Under the Bayesian framework, the definition of Bayes Factor (BF) (see chapter 2.3) can also quantitatively compare the goodness of fitting for two models with different numbers of parameters, which is applied to a wide range of problems [159, 160, 161]. So in the field of GW data analysis, especially for CBC signals, Bayesian Inference methods are heavily used to perform the parameter estimation [41, 76, 75].

We quantify the probability as the belief of certain conclusion to be valid. Such belief should be a real measure between 0 and 1, with 0 meaning definitely false and 1 implying definite truth [156, 157].

We define that the probability of an event  $A$  will happen to be  $P(A)$ . If the event  $A$  does not happen, then we denote that as  $\bar{A}$ , and it is natural in simple situations to expect that event  $A$  either happens or does not happen, which translates to

$$P(A) + P(\bar{A}) = 1. \quad (2.11)$$

This is recognised as the sum rule. We further define the conditional probability as follows: under the condition when the event  $B$  is valid, the probability for event  $A$  to happen is  $P(A|B)$ . It is easy to prove that if  $A$  and  $B$  are independent, then  $P(A|B) = P(A)$  [156, 157].

The joint probability of event  $A$  and  $B$  is written as  $P(A, B)$ , which represents the probability of event  $A$  and  $B$  both being valid. It can be re-written as product of the probability of event  $A$ , and the conditional probability of event  $B$  under the condition that  $A$  valid, namely

$$P(A, B) = P(A)P(B|A). \quad (2.12)$$

This expression is called the product rule [156].

One can notice that in such a factorisation, events  $A$  and  $B$  are equivalent, and so this joint probability could equally be rewritten as

$$P(A, B) = P(B)P(A|B). \quad (2.13)$$

By linking equations 2.12 and 2.13 together, we can obtain an equation

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}. \quad (2.14)$$

Equation 2.14 is recognised as Bayes' theorem named after Thomas Bayes, and it relies on two simple assertions, namely the sum rule (equation 2.11) and product rule (equation 2.12). The proof of these rules are non-trivial, one can find an example from appendix B of [157]. However, the derivation based on them is relatively straightforward.

What makes the Bayesian Inference approach different from the frequentist is not the formulation of equation 2.14, but rather the interpretation of the equation. In both cases, we construct models to explain the data, for example, if we know the mean  $\mu$  and standard deviation  $\sigma$  of a distribution, we can guess it to be Gaussian distributed  $\mathcal{N}(\mu, \sigma)$ .

For practical purposes, we can replace the symbol  $A$  by the data  $D$ , while  $B$  is replaced by the parameter values  $\theta$ , like  $\mu$  and  $\sigma$  in the case of a Gaussian distribution. Such relevant information is represented by  $I$ , which includes but is not limited to the model one chooses [156, 157].

$$P(\theta|D, I) = \frac{P(\theta|I)P(D|\theta, I)}{P(D|I)}. \quad (2.15)$$

1. The *posterior*  $P(\theta|D, I)$  is the degree of belief for  $\theta$ , given the observation of data  $D$  and the underlying information  $I$ .
2. The *prior*  $P(\theta|I)$  is the degree of belief for  $\theta$  under the given prior information  $I$ , before the actual observation.
3. The *likelihood*  $P(D|\theta, I)$  is the probability to obtain the observed data  $D$  if the actual parameter is  $\theta$ , under the information  $I$ .
4. The *evidence* or *marginal likelihood*  $P(D|I)$  is the degree of belief that the observed data are consistent with the underlying information  $I$ . This expression is independent of the  $\theta$  and plays the role of a normalisation constant.

Bayes' Theorem can be interpreted as follows: the degree of belief for  $\theta$ , or the posterior, is updated by the observation  $D$  through the likelihood, from the original prior. The likelihood is not guaranteed to be normalised over  $\theta$  (although by definition it's guaranteed to be normalised over  $D$ ). So the evidence is needed to make sure the posterior would be normalised over  $\theta$ .

Notice that for frequentists, it's meaningless to ask what is the probability of  $\theta$ , because in the physical world, there is only one (unknown) true value of  $\theta$ .

Bayesian Inference is sometimes criticised for the subjectiveness in the choice of the prior. Although one can try to adopt an uninformative prior, it is not a trivial problem to make the prior truly uninformative and such a prior might even not exist [155]. However, as long as the data is of high enough quality, the importance of the prior will be diluted.[156, 157]. The updating property of Bayesian Inference makes it possible to combine information from different sources (for example [103]); meanwhile the definition of probability is more intuitive, makes the result more straightforward, while one should interpret frequentist's results properly before drawing any solid conclusion [153, 154].

## 2.3 The Bayes Factor and Model Selection

In practice, there are frequently situations when different models are to be compared, so it is interesting to know which model is better supported by the model (for example [162]). However, it is not an easy task to compare models quantitatively if the two models have different numbers of parameters. For example, independent and identically distributed (i.i.d.)

Gaussian noise, can be fitted with only a small number of parameters. However, a polynomial fit with a large enough number of terms could always give a nearly perfect result. One can find more scientifically interesting examples from [156] and [157]. It's easy to discard the perfect fitting scheme, since it makes a very strong assumption about the model. As the philosopher William of Occam proposed, one should not add more assumptions if the data could be explained by simpler theory, this spirit of “shaving” complicating components is concluded as the *Occam's razor* which prefers the simpler model when two models can explain the data equally well [156, 157]. What's tricky is the intermediate situation, where a slightly more complicated model could explain the observed data better than the pure noise.

Suppose there are two different models  $M_A$  and  $M_B$ . The dilemma can be expressed as the degree of belief for each model, given the fact that we observed the data  $D$ . This can be formulated as the posterior  $P(M_A|D, I)$

$$P(M_A|D, I) = \frac{P(M_A|I)P(D|M_A, I)}{P(D|I)}. \quad (2.16)$$

The *odds ratio*  $O_{AB}$  is defined as the ratio of two posteriors of model  $M_A$  and  $M_B$ , namely

$$O_{AB} = \frac{P(M_A|D, I)}{P(M_B|D, I)} = \frac{P(M_A|I)}{P(M_B|I)} \times \frac{P(D|M_A, I)}{P(D|M_B, I)} \quad (2.17)$$

Here the ratio  $B_{AB} \equiv \frac{P(D|M_A, I)}{P(D|M_B, I)}$  is recognised as the *Bayes factor*, which is the ratio between the evidences of two models [156, 157].

If a model  $M$  has a parameter  $\theta$ , one needs to marginalise over all possible values of  $\theta$  to compute the evidence

$$\begin{aligned} P(D|M, I) &= \int d\theta P(D, \theta|M, I) \\ &= \int d\theta P(D|\theta, M, I)P(\theta|M, I). \end{aligned} \quad (2.18)$$

If we have no knowledge beforehand, we can assume total ignorance of these two models, thus no preference for either, so the ratio  $\frac{P(M_A|I)}{P(M_B|I)}$  can be set to unity. In such case the odds ratio is completely determined by the Bayes factor [156, 157].

To understand why the Bayes factor naturally disfavours a complicated model, a heuristic proof is given in [157] and [156]. Let us consider a model  $M_A$ , which contains one parameter  $\theta$ , with a flat prior between  $\theta_{min}$  and  $\theta_{max}$ , the prior for any  $\theta$  between the allowable region would be  $(\Delta\theta)^{-1}$  where  $\Delta\theta \equiv \theta_{max} - \theta_{min}$  [156]. Meanwhile the model  $M_B$  contains no

free parameter at all. Thus we can re-express equation 2.18 as

$$P(D|M_A, I) = \frac{1}{\Delta\theta} \int d\theta P(D|\theta, M_A, I).$$

We can replace the integral with an approximation given by the product of the maximum likelihood  $P(D|\hat{\theta}, M_A, I)$  and the effective width  $\delta\theta$  under the definition that

$$\delta\theta \equiv \frac{\int d\theta P(D|\theta, M_A, I)}{P(D|\hat{\theta}, M_A, I)} \quad (2.19)$$

so that  $\delta\theta$  represents the characteristic scale, and further simplify equation 2.18 to

$$\begin{aligned} P(D|M_A, I) &= \frac{1}{\Delta\theta} P(D|\hat{\theta}, M_A, I) \times \delta\theta \\ &= \frac{\delta\theta}{\Delta\theta} P(D|\hat{\theta}, M_A, I). \end{aligned} \quad (2.20)$$

Notice that the integral over  $\theta$  won't be normalised in equation 2.18, as  $\theta$  appears as a condition in the equation. Recall that the normalisation requirement of likelihood applies to the data, rather than the parameter, so it does not generally normalise over  $\theta$  (*c.f.* chapter 2.2).

Normally, with more parameters, one can fit the data better, thus the maximum likelihood  $P(D|\hat{\theta}, M_A, I)$  should get bigger than the maximum likelihood of the simpler model  $M_B$ , or  $P(D|M_B, I)$ . However, the inclusion of every extra parameter will introduce another factor of  $\frac{\delta\theta}{\Delta\theta}$ . So unless the likelihood function has a wide spread over the additional  $\theta$  (namely, a not-so-small  $\delta\theta$  value) or a much higher likelihood value, or a combination of both, the simpler  $M_B$  will be preferred. We can easily see that this formalism will punish the fine-tuned model, while supporting the robust model which can fit data well with less sensitivity to the change of parameter.

The odds ratio, or practically, the Bayes factor tells how much the data favours one model against another. In practice, we use the table 2.1 as a guide line to interpret the Bayes factor into strength of evidence [163].

The odds ratio provides a quantitative Occam's razor, which can compare models with different dimensionality, provides a practical rule to execute Occam's principle "entities must not be multiplied beyond necessity" [156, 157, 164]. Thus it is of primary concern to develop techniques to compute the odds ratio (for example [165, 166, 167, 168, 160, 169] and reference within), in order to apply this method to astrophysical problems [161, 113, 12, 170, 101, 171].

$O_{AB}$	strength of evidence
$< 1$	Negative (support $M_B$ )
$1 < O_{AB} < 10^{1/2}$	Negligible support of $M_A$
$10^{1/2} < O_{AB} < 10$	Substantial support of $M_A$
$10 < O_{AB} < 10^{3/2}$	Strong support of $M_A$
$10^{3/2} < O_{AB} < 100$	Very strong support of $M_A$
$100 < O_{AB}$	Decisive support of $M_A$

Table 2.1: odds ratio value and their corresponding strength of evidence. This table serves just as a reference purpose. Table adapted from [163].

## 2.4 Stochastic Samplers

The methodology of frequentist statistics once dominated the field of probability for a good reason [156]. In order to perform the Bayesian Inference, the posterior distribution over the parameter space needs to be computed. If this can not be achieved analytically, we have to compute numerically, which may involve a large amount of computing resources. On the other hand, the frequentist probability seems to provide a natural tool for distinguishing the two models, namely the *p-value*, defined as the estimated probability of obtaining the observed result, under the assumption that the null hypothesis  $H_0$  is true. [172]. However, one need to notice that the definition of the p-value is complicated and a large portion of users misuse it in this context [172]. This is not saying that frequentist method is wrong [154, 153], but simply implies the fact that the correct interpretation of the result requires non-trivial effort.

A naive way of performing Bayesian inference is by placing grid points uniformly over the parameter space [173, 174, 175, 176]. Placing only 10 points in each dimension means  $10^N$  points in total for  $N$  dimensions. In such cases, the computational burden will grow exponentially with the dimensionality of the underlying information, which shortly makes the strategy unaffordable. There are cleverly developed methods to bypass such a curse of dimensionality, among them, the two most recognised and widely applied methods in the field of gravitational wave astronomy are Markov Chain Monte Carlo (MCMC) (see for example [156, 177, 178, 179, 180] and the references within) and nested sampling (see for example [157, 181, 182] and the references within).

### 2.4.1 Markov Chain Monte Carlo (MCMC)

A Markov Chain is a process where the next point's position depends only on the previous point's position, while Monte Carlo is a general methodology which adopts random numbers to perform computations [156].

The most common and simple MCMC algorithm is the Metropolis-Hastings algorithm [156,

183, 184]. To start with, a random location is selected in the parameter space. For the  $i + 1^{th}$  step, the location  $\theta_{i+1}$  is determined by the previous  $i^{th}$  step  $\theta_i$  by the *transition probability* or *transition kernel*  $P(\theta_{i+1}|\theta_i)$  [156]. A convenient setup includes a *proposal density*  $q(\theta^*|\theta_i)$  and a corresponding *acceptance probability*. A proposal is randomly drawn from the probability distribution function (PDF)  $q(\theta^*|\theta_i)$ , and it would be accepted with the *acceptance probability*, or otherwise, rejected, in which case  $\theta_{i+1} = \theta_i$ . One can prove that the sampler's density will be representative of the posterior PDF [156].

To be explicit, the new point's posterior is calculated and then compared with the previous posterior, forming a *Metropolis ratio*  $r$ ,

$$r = \frac{P(\theta^*|D, I) q(\theta_i|\theta^*)}{P(\theta_i|D, I) q(\theta^*|\theta_i)}. \quad (2.21)$$

The new point is accepted with an *acceptance probability*  $\alpha$  defined as [156]

$$\alpha(\theta_i, \theta^*) \equiv \min(1, r) = \min \left( 1, \frac{P(\theta^*|D, I) q(\theta_i|\theta^*)}{P(\theta_i|D, I) q(\theta^*|\theta_i)} \right) \quad (2.22)$$

Under such a design, an MCMC sampler can start from anywhere in the parameter space that is allowed by the prior. Since the starting point is more likely to be located in regions with relatively low posterior, the following points will generally be going uphill in the posterior, but sometimes it is also possible for the sampler to jump downwards. After a period of *burn-in* which could take as much as half of the samples, the memory of the start point would be washed out. In order to prevent the influence of the choice of starting point, the burn-in stage would generally be discarded, so that the remaining samples will be representative of the posterior [185, 156, 177].

The algorithm of MCMC is sampling towards the target distribution  $P(\theta|D, I)$ , which is call the *stationary distribution* of the Markov chain [156]. An MCMC chain should keep sampling until convergence to a stationary distribution is achieved [185]. In order to achieve convergence, the sampler must satisfy three properties [186, 156]:

### 1. Irreducible

The sampler must have positive probability of sampling in points with non-zero posterior, instead of becoming stuck in one specific region.

### 2. Aperiodic

The sampler should NOT oscillate between finite states.

### 3. Positive recurrent

There exists a stationary distribution, otherwise it is meaningless to discuss the convergence problem.

As the Metropolis-Hastings sampler satisfies all requirements for regular posteriors, it is guaranteed to converge to a stationary distribution.

The criterion of *detailed balance* is a sufficient but not necessary condition for the convergence of MCMC, illustrated as follows

$$P(\theta_i)q(\theta_{i-1}|\theta_i) = P(\theta_{i-1})q(\theta_i|\theta_{i-1}) \quad (2.23)$$

The concept of detailed balance in thermodynamics could help to understand this requirement. In thermodynamics, if we define the probability of a particle to be in state  $\theta_i$  as  $P(\theta_i)$ , and the probability of it to jump from state  $\theta_i$  to state  $\theta_{i-1}$  as  $q(\theta_{i-1}|\theta_i)$ , detailed balance requires that during a certain period, the probability for a particle to jump from state  $\theta_i$  to state  $\theta_{i-1}$  should be exactly the same probability as for it to jump from state  $\theta_{i-1}$  to state  $\theta_i$  [187].

Since detailed balance is a stronger requirement than convergence, some Markovian Chain that does not meet the requirement of detailed balance can still be able to converge to the target distribution (for example, [188]). However, it is much easier to check and to implement if one has chosen the criterion to be detailed balance. We can prove that detailed balance can lead to the convergence on the target distribution as follows [156],

$$\begin{aligned} \text{Joint probability}(\theta_i, \theta_{i+1}) &= P(\theta_i)P(\theta_{i+1}|\theta_i) \\ &= P(\theta_i)q(\theta_{i+1}|\theta_i)\alpha(\theta_i, \theta_{i+1}) \\ &= P(\theta_i)q(\theta_{i+1}|\theta_i) \min \left( 1, \frac{P(\theta_{i+1})q(\theta_i|\theta_{i+1})}{P(\theta_i)q(\theta_{i+1}|\theta_i)} \right) \\ &= \min (P(\theta_i)q(\theta_{i+1}|\theta_i), P(\theta_{i+1})q(\theta_i|\theta_{i+1})) \\ &= \alpha(\theta_{i+1}, \theta_i)P(\theta_{i+1})q(\theta_i|\theta_{i+1}) \\ &= P(\theta_{i+1})P(\theta_i|\theta_{i+1}) \end{aligned} \quad (2.24)$$

If we marginalise over all possible  $\theta_i$  values, then

$$\begin{aligned} \int d\theta_i P(\theta_{i+1}|\theta_i)P(\theta_i) &= \int d\theta_i P(\theta_i|\theta_{i+1})P(\theta_{i+1}) \\ &= P(\theta_{i+1}) \int d\theta_i P(\theta_i|\theta_{i+1}) \\ &= P(\theta_{i+1}) \end{aligned} \quad (2.25)$$

where in the second line, the integral is naturally normalised to unity [156].

Equation 2.25 guarantees that the probability of sampling any parameter  $\theta$  is just the posterior of that parameter. This property makes it very easy to infer any numerical integral concerning the probability density. For any function  $f(\theta)$ , the integral over parameter space can be



approximated by numerical summation

$$\int d\theta f(\theta)P(\theta) \approx \frac{1}{N} \sum_{i=1}^N f(\theta_i) \quad (2.26)$$

### 2.4.2 Nested Sampling

Although MCMC is very versatile and is able to perform Bayesian Inference in a large variety of problems, especially problems with high dimensionality, it has its own shortcoming. For example, most of the MCMC samples will concentrate on the high posterior regions, and are unable to explore the vast majority of the parameter space. This is helpful to keep the MCMC focused in interesting regions, thus alleviating the so-called *curse of dimensionality*, which is the outstanding problem that the computational requirements scale faster than the dimensionality. However, this also means that the part of parameter space with low posterior will be mostly ignored, and it'll become a big problem for normal MCMC samplers to estimate the evidence accurately (although notably there have been some effort to alleviate this problem, see for examples [189, 165]). The Nested Sampling method, on the other hand, is designed to estimate the evidence, in addition having the ability to perform parameter estimation [157, 166, 167].

To start with, let's look at the calculation of the canonical partition function in statistical mechanics. The partition function  $Z(\beta)$  is defined as

$$Z(\beta) = \int d\theta \exp(-\beta E(\theta)). \quad (2.27)$$

If one has the knowledge of the Density of States (DOS)

$$g(E) = \int d\theta \delta(E - E(\theta)), \quad (2.28)$$

where  $\delta$  is the Dirac delta function, one can transform equation 2.27 into

$$Z(\beta) = \int dE g(E) \exp(-\beta E). \quad (2.29)$$

Now let us step back to the Bayesian point of view, and we regard  $E(\theta)$  as the  $-\log P(D|\theta, M, I)$ , and

$$g(E) = \int d\theta P(\theta|M, I) \delta(E - E(\theta)).$$

We can regard this as assigning a measure according to the prior. In this form, the evidence value is simply the partition function when  $\beta = 1$  [190].

The philosophy of Nested Sampling is to achieve the integral  $Z = \int_{-\infty}^{\infty} dE g(E) \exp(-E)$  numerically.

We define *prior mass* as

$$X(L) = \int_{P(D|\theta, M, I) > L} d\theta P(\theta|M, I) \quad (2.30)$$

which is integrated over the parameter space that contains a likelihood value larger than set  $L$ . The  $X$  value is constrained between 0 and 1 [157, 167].

The evidence integral can be translated as

$$\begin{aligned} Z &= \int_{-\infty}^{\infty} dE g(E) \exp(-E) \\ &= \int_0^1 dX L(X) \\ &\approx \sum_{i=2}^N L_i (X_i - X_{i-1}). \end{aligned} \quad (2.31)$$

The algorithm of Nested Sampling will sample a number of points  $\theta_i$  from the prior, and calculate the corresponding likelihood  $L_i$ , which is later sorted as  $L_1 < L_2 < \dots < L_N$ . Naturally, one can conclude  $X_1 > X_2 > \dots > X_N$ , as illustrated in figure 2.1 [157].

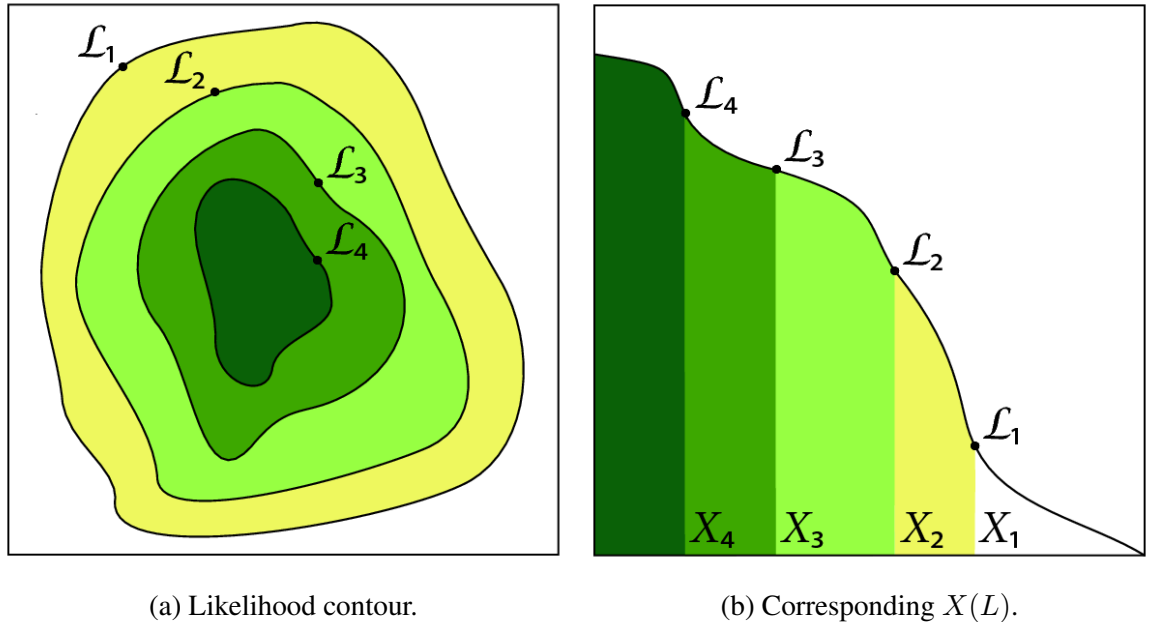


Figure 2.1: An illustration of likelihood contours and corresponding  $X(L)$ . Since the definition of  $X(L)$  is the prior mass that's integrated over the parameter space with likelihood larger than  $L$ , the larger the likelihood value, the smaller the  $X$  value is. Picture reproduced from [191].

To set up the Nested Sampling algorithm, one needs to spread a number  $n$  of (for example

$n \sim 1,000$ ) live points uniformly in the prior. The point with the smallest likelihood value will be recorded as the first sampled point  $\theta_1$ . That live point would be updated by a randomly drawn point, which will have a likelihood value bigger than it. In this way, the sampler will go upwards to higher likelihood regions, just like stripping off layers of an onion. For each update at step  $k$ , the mass  $X_k$  is defined as the mass  $X(L)$  with the lowest likelihood among all live points. For a random update, the shrink rate  $t = X_{k+1}/X_k$  follows a distribution

$$P(t) = nt^{n-1}$$

thus gives

$$\langle \log t \rangle = (-1 \pm 1)/n$$

Since  $X_0 = 1$ , while  $X_k = \prod_{j=1}^k t_j$ , one can prove that

$$\langle \log X_k \rangle = (-k \pm \sqrt{k})/n \quad (2.32)$$

Combined with equation 2.31, one can explicitly calculate the estimation of evidence. One can also compute the information  $\mathcal{H}$  defined as

$$\mathcal{H} = \int dX P(X) \log [P(X)] \approx \sum_k \frac{L_k (X_k - X_{k-1})}{Z} \log \left[ \frac{L_k}{Z} \right]. \quad (2.33)$$

From the information  $\mathcal{H}$ , one can estimate the uncertainty in the estimation of evidence, so that

$$\log Z \approx \log \left( \sum_k L_k (X_k - X_{k-1}) \right) \pm \sqrt{\frac{\mathcal{H}}{n}}. \quad (2.34)$$

The algorithm is designed to have sparse sampling in low likelihood regions, where the contribution to the evidence is negligible, while high likelihood regions would be densely sampled so that the uncertainty in the final estimation can be minimised [157, 166, 167].

The algorithms of MCMC and Nested Sampling listed in chapter 2.4 are just the rules of thumb. In order to deal with specific problems, there are lots of different variants to implement the idea of both MCMC and Nested Sampling, so that the methods are tailored to different requirements. In the remainder of this chapter, the main variants of MCMC and Nested Sampling will be discussed.

### 2.4.3 Parallel Tempering MCMC

One of the key problems for MCMC is the multi-modal problem, where there exist multiple modes in the posterior distribution, or multiple distinct regions within the parameter space

that can explain the data with high confidence. For a simple Metropolis-Hastings algorithm, if the sampler wants to jump from one mode to another, it has to go through a ‘valley’ in the posterior, which usually takes a sequence of low probability downwards jumps, which the sampler will take an extremely long time to achieve, so although theoretically it is possible, it is practically exceedingly unwise.

The philosophy of parallel tempering MCMC is to implement multiple chains, with some chains sampling a flatter posterior, and the others sampling the target distribution [192, 193, 194]. With the communication between the chains enabled, it can guide the samplers to achieve jumps between distinct regions of the parameter space.

The flattened distribution has a posterior

$$P_T(\theta|D, I) \propto P(\theta|I)P(D|\theta, I)^\beta$$

where the parameter  $T \equiv \frac{1}{\beta}$  is defined as the “temperature” of a chain. In the scheme of parallel tempering, a temperature ladder is constructed, so a series of (say,  $n$ ) chains are configured with different temperatures,  $T_0 < T_1 < \dots < T_n$ , where the lowest temperature  $T_0 = 1$ . It is not hard to see that higher temperature chains are more similar to the prior, and the posterior with temperature of  $\infty$  degenerates to the prior [192, 193, 194, 165, 156, 195, 177].

Neighbouring chains have a constant ratio in temperature. For a chain with higher temperature, the posterior is flatter, and it is more likely for the sampler to jump from one mode to the other. Occasionally, a swap between two neighbouring chains labelled as  $j$  and  $j + 1$  is proposed, the acceptance of such a swap being given by the probability

$$r = \min \left\{ 1, \frac{P_{T_j}(\theta_{i+1}|D, I)P_{T_{j+1}}(\theta_i|D, I)}{P_{T_j}(\theta_i|D, I)P_{T_{j+1}}(\theta_{i+1}|D, I)} \right\}. \quad (2.35)$$

From this scheme, one can keep the low temperature chains sampling according to the target distribution, while the high temperature chains will be auxiliary in the sense of guiding swaps between modes [177, 195].

#### 2.4.4 Delayed Rejection MCMC

For most cases, the bottleneck of computation for an MCMC algorithm lies in the numerical calculation of the posterior. So it is somehow a waste of computational ability when the proposed point  $\theta_{(1)}^*$  is finally rejected, according to the acceptance ratio. We slightly modify equation 2.22 into

$$\alpha_1(\theta_i, \theta_{(1)}^*) = \min \left( 1, \frac{P(\theta_{(1)}^*)}{P(\theta_i)} \frac{q_1(\theta_i|\theta_{(1)}^*)}{q_1(\theta_{(1)}^*|\theta_i)} \right).$$

Although some rejection is necessary to give us information about the posterior, it would be efficient to not discard the calculated information immediately, but reuse it by proposing candidates incorporating the information, with  $q_2(\theta_{(2)}^* | \theta_{(1)}^*, \theta_i)$

The acceptance ratio for the newly proposed point  $\theta_{(2)}^*$  is

$$\alpha_1(\theta_i, \theta_{(1)}^*, \theta_{(2)}^*) = \min \left( 1, \frac{P(\theta_{(2)}^*)}{P(\theta_i)} \frac{q_1(\theta_{(1)}^* | \theta_{(2)}^*)}{q_1(\theta_{(1)}^* | \theta_i)} \frac{q_2(\theta_i | \theta_{(1)}^*, \theta_{(2)}^*)}{q_2(\theta_{(2)}^* | \theta_{(1)}^*, \theta_i)} \frac{1 - \alpha_1(\theta_{(2)}^*, \theta_{(1)}^*)}{1 - \alpha_1(\theta_i, \theta_{(1)}^*)} \right) \quad (2.36)$$

The final term is to account for the fact that the proposed  $\theta_{(1)}^*$  was not accepted. This scheme can be extended to higher order [196, 197, 198].

### 2.4.5 Reversible Jump MCMC

Normal MCMC is relatively powerless in dealing with model selection due to the inaccurate estimation of the evidence (see for example [165, 159] for efforts of computing evidence with MCMC). However, the reversible jump MCMC can propose jumps between two different models. Assume two models with different dimensions, in addition to the usual Metropolis-Hasting algorithm, one should also design a trans-dimensional proposal density, which would be in equilibrium under movements between the models [160, 197].

The central problem of reversible jump MCMC is the construction of the jump proposal between different models. A badly constructed jump proposal between models might lead to an inefficient swap between models, and might lead to biased results, so attempts have been made to construct robust jump proposals between models [169].

### 2.4.6 Automated MCMC

For a Markov chain, the points are expected to be correlated with the adjacent points; the closer the points are, the more correlated they are. The correlation between a point and another point separated by  $\tau$  is represented as *autocorrelation function*

$$R(\tau) = \frac{\mathbb{E}[(\theta(i) - \mu)(\theta(i + \tau) - \mu)]}{\sigma^2}, \quad (2.37)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation for the distribution, and the average is taken over index  $i$ . We also identify the characteristic length of the autocorrelation function as the *autocorrelation length*.

When optimising the MCMC algorithm, it is not trivial to make sure the proposal density is well designed. If the proposed jump is too large, the acceptance ratio could be too low to be efficient; on the other hand, if the proposal jump is too small, then the sampled points

will have a very high autocorrelation length, so the MCMC will take a long time until it explores a large enough portion of the parameter space [199, 156]. An optimal proposal will lie in somewhere in the middle, enabling a high enough acceptance probability, while the autocorrelation length is low enough so that the sampling process will be more efficient.

Efforts have been made to make this process automatic and robust. For example, Gregory proposed a method [156, 200] that before running the MCMC algorithm, a target *acceptance rate* is set. From time to time, the actual acceptance rate will be compared with the reference [177]. The similar idea is implemented in the MCMC version of LALInference within the LSC Algorithm Library (LAL), however, by proposing new candidates from the covariance matrix of sampled points, which makes the automation process more versatile [41, 76].

### 2.4.7 Affine Invariant MCMC

Recently, a specific implementation of MCMC has attracted huge attention [201]. By designing an sampling algorithm invariant to affine transformation, the sampler should be equally efficient over all convex bodies for given dimensionality [201].

In the design of such algorithm, there exist a group of copies of the system each with a sampler, or an *ensemble* of *walkers*, and for step  $t$  every walker  $X_k$  is evolved with

$$X_k(t) \rightarrow Y = X_j + Z(X_k(t) - X_j) \quad (2.38)$$

where  $X_j$  with  $j \neq k$  is another walker in the ensemble, and  $Z$  is a scaling variable that makes sure the detailed balance is satisfied.

The most beneficial aspect of this proposal is that it could achieve a low autocorrelation length with a relatively high acceptance rate [201]. So in principle this method could be very efficient. Also, normal implementation of MCMC usually requires a tuning of the running parameters to make it work as expected. In this affine invariant MCMC, since the proposal is independent of the covariance matrix, such tuning is not necessary any longer, making it versatile and convenient to use. The EMCEE python implementation of it has become one of the most popular MCMC tools recently [202]. And in the field of GW data analysis, the idea is customised to perform efficient PE [111].

### 2.4.8 MultiNest

There are also different variants for the implementation of Nested Sampling. As the update of the live points should be in principle randomised according to prior mass, which means the live points will become more and more concentrated, until they shrink into a small region

in the parameter space [157, 166, 167], some proposal methods of the probability could be computationally expensive [203].

The MultiNest algorithm use ellipsoids to approximate the distribution, so that it can efficiently propose new points from within the ellipsoidal approximation. So this method is especially efficient for multi-modal scenarios [168, 181, 191].

### 2.4.9 Combinations

The methods mentioned in this section can be combined together to construct new methods [186]. For example one can combine the delayed rejection and reversible jump [197], combine MCMC and Nested Sampling [203] or combine importance sampling with MultiNest [191]. Importance sampling is a strategy to propose independent samples from a simpler form of distribution than the target distribution, thus one can perform numerical integral efficiently. Notice also that [111] implemented an affine-invariant ensemble sampler together with parallel tempering to tailor for the multi-modal feature, and efforts are being taken to implement this method in LAL to achieve fast analysis.

## Chapter 3

# Bayesian Model Selection on Anti-Glitch Models

The application of Bayesian Inference is vital to problems which involve model selection. There are situations where multiple models can explain the observed data comparably well, and with similar complexity, or cases where more complex models explain the data better, then the general Occam's razor will not be practical. As we've discussed in chapter 2.3, Bayesian Inference methods naturally provide a quantitative Occam's razor. In this chapter, an actual application of Bayesian Model Selection is performed on such a problem from the observation of an anti-glitch event [162].

Neutron stars generally have strong magnetic fields and rotate with a very steady period. A phenomenon known as a glitch has been observed in multiple pulsars, where the pulsar undergoes a sudden frequency change, and usually an increase of frequency is observed. It has been suggested that glitches are caused by internal processes of the pulsar [204]. With the interior of the neutron star being in a superfluid state, the magnetic fields within the core are concentrated into quantised vortex lines, and the vortex can be pinned to the crust, which causes the surface crust rotate slightly slower than the inner core. A sudden twist would lead to the spin up of the crust, which leads to the observed glitch [205].

However, there's a small group of neutron stars with exceedingly high magnetic field ( $\sim 10^{14}G$ ). Such strong magnetic fields can support some interesting phenomena, like soft gamma repeaters [206]. In previous work, an unusual spin-down glitch, or anti-glitch is also observed in a magnetar [162].

In the year 2013, *Archibald et al.* [162] discovered an unexpected anti-glitch phenomenon in magnetar 1E 2259+586. Unlike a normal glitch, which undergoes a sudden spin up, this magnetar experienced a sudden spin-down. The mechanism which caused this phenomenon is still under discussion (e.g. [207, 208, 209, 210, 211]), but to our knowledge no model



explicitly predicted an anti-glitch prior to this discovery, although [205] predict a similar potential phenomenon in SGR 1900+14 [162].

The data analysis performed by [162] shows that during the observation, 1E 2259+586 undergoes two timing events separated by 50–90 days. The first event is a certain anti-glitch, while the nature of the second event is less certain. If it is also an anti-glitch this might require a qualitatively different physical model to explain its origin. Importantly, however, the analysis performed in [162] was unable to distinguish between these two types of glitch for the second event.

Since we know very little about the mechanism behind such a rare phenomenon, any information about it could be helpful to understand its physical cause. In this work we seek to use the data themselves, employing the methods of Bayesian model selection, to distinguish between two competing models, wherein the second event is a glitch, or anti-glitch, respectively. More specifically, we compute the ratio of the *evidence* for each model (as defined in Section 2.3) and investigate whether this ratio favours one model over the other.

The structure of this chapter is as follows. In Section 3.1 we briefly review relevant details of the model for the time of arrival of pulses from the progenitor. Section 3.2 presents the results of our analysis, including a careful check on their robustness. Finally Section 3.3 summarises our conclusions.

## 3.1 Timing Model

The magnetar 1E 2259+586 was routinely observed by the X-ray Telescope (XRT) onboard Swift every 2-3 weeks, with more frequent observations being made shortly after discovering the first anti-glitch event reported in [162]. The observations give the time of arrival (TOA) of each X-ray pulse (which can be corrected to the solar system barycenter), which in turn gives the pulse phase of the magnetar. Together with each TOA, the X-ray flux is also recorded. An increase in the X-ray flux helps to pinpoint the epoch of the first glitch event, while for the second event, no obvious flux change was detected – thus contributing to the confusion about the second event’s type. The observational data of the magnetar is shown as in figure 3.1 (reproduced from [162]).

We model the magnetar’s phase evolution,  $\phi(t)$ , with the standard Taylor expansion of frequency and frequency derivatives [212] using terms up to second order, when no (anti-)glitch happens, it will follow that

$$\begin{aligned}\phi(t) &= \int_{t_0}^t \nu(t') dt' \\ &= \int_{t_0}^t dt' [\nu_0 + \dot{\nu}t' + \mathcal{O}(\ddot{\nu}t'^2)] \\ &= \nu_0(t - t_0) + \frac{1}{2}\dot{\nu}(t - t_0)^2 - \Delta\phi(t).\end{aligned}\tag{3.1}$$

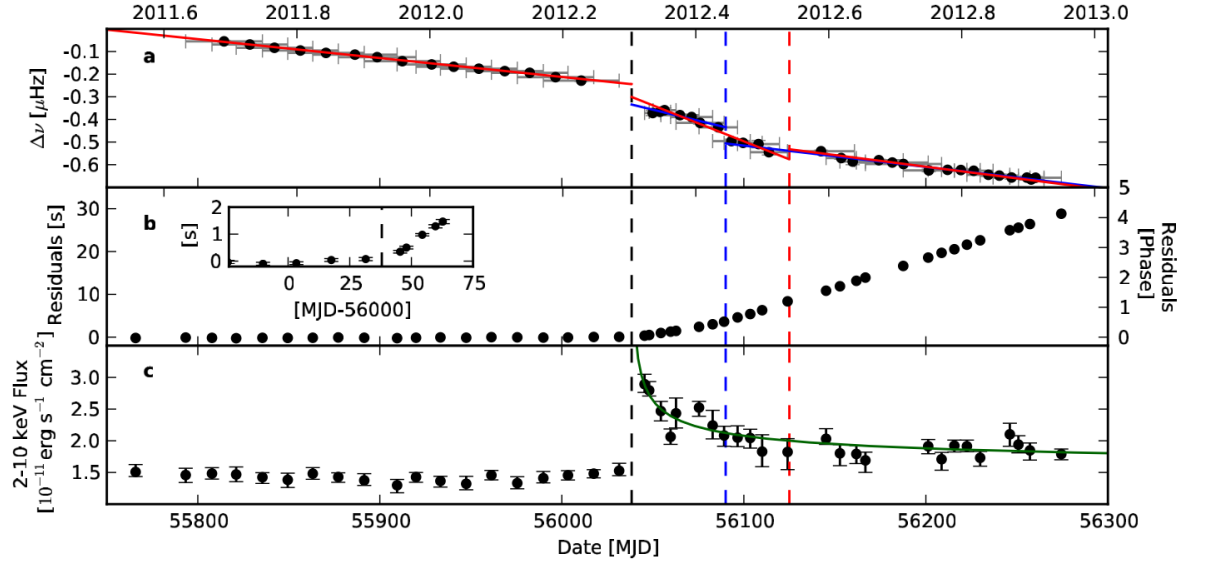


Figure 3.1: Upper panel: 1E2259+586s spin frequency as a function of time, determined by short-term fitting of (typically) five TOAs. The grey horizontal error bars indicate the ranges of dates used to fit the frequency, and the vertical error bars (generally smaller than the points) are standard  $1\sigma$  uncertainties. The red and blue solid lines in a represent the fits to the pulse TOAs, with red represents an anti-/normal glitch pair model, and the blue represents a successive anti-glitch model. Middle panel: Timing residuals (differences between the initial model and observed data) of 1E2259+586 after fitting only for the pre-anti-glitch timing solution. The inset shows the same timing residuals, zooming in on the anti-glitch epoch. Lower panel: The absorbed 2 – 10keV X-ray flux. The error bars indicate the  $1\sigma$  uncertainties, and the green line is the best-fit power-law decay curve with an index of  $-0.38 \pm 0.04$ . The dashed vertical lines running through all panels indicate the glitch epochs, the black line being the anti-glitch, and blue and red lines the second event in the models of the best fit. The timing residuals for these fits can be seen in the Supplementary Information. Figure reproduced from [162]

Notice that the period of a pulsar's phase is unity instead of  $2\pi$ , so for each TOA  $t_i$ , there is a integer  $N_i$  so that the pulsar has been through  $N_i$  cycles since  $t_0$ . If we define  $\phi(t_0) = 0$ , then  $\phi(t_i) = N_i$ , and we can write

$$\Delta\phi(t_i) = \nu_0(t_i - t_0) + \frac{1}{2}\dot{\nu}(t_i - t_0)^2 - N_i. \quad (3.2)$$

The effect of a (anti-)glitch on the phase timing model will be a sudden change in the frequency and frequency evolution after the event [213]. Suppose that the epoch of the glitch is  $T_g$ , then the accumulated phase at  $T_g$  is

$$\phi_g = \nu_0(T_g - t_0) + \frac{1}{2}\dot{\nu}(T_g - t_0)^2 \quad (3.3)$$

meanwhile the instantaneous frequency  $\nu_g$  is

$$\nu_g = \nu_0 + \dot{\nu}(T_g - t_0)$$

The glitch will cause a sudden change in all high order terms. We take the sudden change of frequency and its derivative as  $\Delta\nu$  and  $\Delta\dot{\nu}$  separately. The frequency evolution after the glitch is then written as

$$\nu(t) = (\nu_g + \Delta\nu) + (\dot{\nu} + \Delta\dot{\nu})(t - T_g) + \mathcal{O}\ddot{\nu}(t - T_g)^2$$

for  $t > T_g$ . And then we can write the time evolution of phase by

$$\phi(t) = \int_{T_g}^t \nu(t')dt' + \phi_g + \Delta\phi_g \quad t > T_g. \quad (3.4)$$

Therefore the difference between the phase model with and without the (anti-)glitch under the assumption that no decaying frequency increment is required due to none being fitted by [162] is

$$\Delta\phi(t) = \Delta\nu(t - T_g) + \frac{1}{2}\Delta\dot{\nu}(t - T_g)^2 + \Delta\phi_g \quad (3.5)$$

where  $\Delta\phi_g$  captures the residual change of phase not well modeled by the frequency change.

We further define

$$R_i = \frac{\phi_i - N_i}{\nu}$$

as the time residual after subtracting the model predictions from the data. Here  $\phi_i$  is the predicted pulse phase at the  $i^{th}$  observation time  $t_i$  (i.e.  $\phi(t_i)$ ),  $N_i$  is the exact phase at the TOA, which by definition is an integer, and  $\nu$  is the frequency according to the model.

Together with the observed timing uncertainty,  $\sigma_i$ , we can form

$$\chi^2 = \sum_{i=1}^N \left( \frac{R_i}{\sigma_i} \right)^2. \quad (3.6)$$

For the purpose of performing Bayesian Inference, we define the likelihood as proportional to  $\exp(-\chi^2/2)$ , and use it to evaluate the evidence for each model. We apply the Nested Sampling algorithm [203] (*c.f.* chapter 2.4.2) to compute the evidence of both models, and further the Bayes factor is calculated to determine which model is more supported by the data.

## 3.2 Results

In our analysis, the two models under consideration only differ in the sign of the frequency change for the second event. In order to avoid undue influence of the prior range on the Bayes Factor, we assign identical prior ranges to all common parameters in both models.

### 3.2.1 Setting the Priors

From table 1 of [162], the parameters are given to be  $\nu_0 = 0.143285110 \pm (4 \times 10^{-9})\text{Hz}$ ,  $\dot{\nu} = -9.80 \pm 0.09 \times 10^{-15}\text{Hz s}^{-1}$ . The Epoch (MJD) is the time  $t_0$  when  $\nu(t_0) = \nu_0$ . Since the magnetar has been observed for a long time (e.g. [214, 215]), and the spin before the anti-glitch is not of interest, we fix those parameters to be constants.

In table 3.1, the estimated value and corresponding uncertainties for glitch parameters are given from the analysis of TEMPO2 fitting, adapting from [162]. Notice that the phase parameter  $\Delta\phi_g$  is not given in [162]. It is usually a small value, so we set the mean of  $\Delta\phi_g = 0$  while the uncertainty  $\sigma_{\Delta\phi_g} = 0.05$ . Later we will find out that this  $\sigma$  choice encapsulates the parameter with the highest posterior value.

In both models, there are two independent (anti-)glitch events and for each event there are 4 parameters required to describe it: its epoch  $t$  and the changes in the frequency  $\Delta\nu$ , its first derivative  $\Delta\dot{\nu}$  and the phase  $\Delta\phi_g$  caused by the event. Thus, in total there are eight parameters for each model. In order that the two models should have a common parametrisation we suppose that in the second model, after the second event (which is a normal glitch in this model), the frequency becomes  $\nu_g = \nu - \Delta\nu$  while in the first model, after the second event (which is an anti-glitch) the frequency becomes  $\nu_g = \nu + \Delta\nu$ . In this way  $\Delta\nu$  is a positive parameter for the second event in both models. With this design, the two models can have

Parameter	Value $\theta$	Uncertainty $\sigma_\theta$
Epoch(MJD)	55,380.000	—
$\nu(\text{s}^{-1})$	0.143,285,110	$4 \times 10^{-9}$
$\dot{\nu}(\text{s}^{-2})$	$-9.80 \times 10^{-15}$	$9 \times 10^{-17}$
Anti-/normal glitch pair		
Epoch1(MJD)	56,035	2
$\Delta\nu_1(\text{s}^{-1})$	$-4.5 \times 10^{-8}$	$6 \times 10^{-9}$
$\Delta\dot{\nu}_1(\text{s}^{-2})$	$-2.7 \times 10^{-14}$	$2 \times 10^{-15}$
Epoch2(MJD)	56,125	2
$\Delta\nu_2(\text{s}^{-1})$	$3.6 \times 10^{-8}$	$7 \times 10^{-9}$
$\Delta\dot{\nu}_2(\text{s}^{-2})$	$2.6 \times 10^{-14}$	$2 \times 10^{-15}$
Successive Anti-glitch		
Epoch1(MJD)	56,039	2
$\Delta\nu_1(\text{s}^{-1})$	$-9 \times 10^{-8}$	$1 \times 10^{-8}$
$\Delta\dot{\nu}_1(\text{s}^{-2})$	$-1.3 \times 10^{-14}$	$4 \times 10^{-15}$
Epoch2(MJD)	56,090	3
$\Delta\nu_2(\text{s}^{-1})$	$-6.8 \times 10^{-8}$	$8 \times 10^{-9}$
$\Delta\dot{\nu}_2(\text{s}^{-2})$	$1.1 \times 10^{-14}$	$4 \times 10^{-15}$

Table 3.1: Parameter estimation for the timing model from TEMPO2 fitting, including epoch, frequency and first derivative of frequency before glitch and after the two glitch events for both models. Table modified from [162]. Note this is being used to inform the choice of prior.

exactly the same priors, thus minimising the influence of the choice of prior on the final value of the Bayes Factor.

For the epoch of the first anti-glitch event there is an obvious change in flux between the 19<sup>th</sup> and 20<sup>th</sup> observation; hence we set the prior for the epoch to be flat between these two data points. For the priors on other parameters we make use of their estimated values  $\theta$ , together with their uncertainties  $\sigma_\theta$ , as reported in [162]. Specifically we adopt a conservative, uniform prior of width equal to  $2n$  times the uncertainty for each parameter – where we will adopt different values of  $n$  in order to explore the robustness of our results to the choice of prior, i.e. to check that our prior boundaries contain the vast bulk of the likelihood.

Thus for each parameter (and where  $M_i$  refers to model  $i$ ) the lower boundary of the uniform prior is set to be  $\min(\theta_{M_1} - n\sigma_{\theta;M_1}, \theta_{M_2} - n\sigma_{\theta;M_2})$ , while the upper boundary is set to be  $\max(\theta_{M_1} + n\sigma_{\theta;M_1}, \theta_{M_2} + n\sigma_{\theta;M_2})$ . Note however, that since  $\Delta\nu$  for the second event is always positive, its lower limit is set to be  $\min(0, \Delta\nu_{M_1} - n\sigma_{\Delta\nu;M_1}, \Delta\nu_{M_2} - n\sigma_{\Delta\nu;M_2})$ .

We tested cases for  $n$  equals to 3, 5 and 10, to make sure our conclusion would be independent from any specific choice of prior range. In tables 3.2, 3.3 and 3.4, we show the upper and lower boundaries for cases with  $n$  equals to 3, 5 and 10.

Parameter	Lower Boundary	Upper Boundary
Epoch1(MJD)	56031.484167333	56045.4437822243
$\Delta\nu_1(\text{s}^{-1})$	$-1.2 \times 10^{-7}$	$-2.7 \times 10^{-08}$
$\Delta\dot{\nu}_1(\text{s}^{-2})$	$-3.3 \times 10^{-14}$	$-1 \times 10^{-15}$
$\Delta\phi_{g,1}$	-0.15	0.15
Epoch2(MJD)	56081	56131
$\Delta\nu_2(\text{s}^{-1})$	0	$9.2 \times 10^{-08}$
$\Delta\dot{\nu}_2(\text{s}^{-2})$	$-1 \times 10^{-15}$	$3.2 \times 10^{-14}$
$\Delta\phi_{g,2}$	-0.15	0.15

Table 3.2: Upper and lower boundaries of timing parameters in the case where  $n = 3$ .

Parameter	Lower Boundary	Upper Boundary
Epoch1(MJD)	56031.484167333	56045.4437822243
$\Delta\nu_1(\text{s}^{-1})$	$-1.4 \times 10^{-7}$	$-1.5 \times 10^{-08}$
$\Delta\dot{\nu}_1(\text{s}^{-2})$	$-3.7 \times 10^{-14}$	$7 \times 10^{-15}$
$\Delta\phi_{g,1}$	-0.25	0.25
Epoch2(MJD)	56075	56135
$\Delta\nu_2(\text{s}^{-1})$	0	$1.08 \times 10^{-07}$
$\Delta\dot{\nu}_2(\text{s}^{-2})$	$-9 \times 10^{-15}$	$3.6 \times 10^{-14}$
$\Delta\phi_{g,2}$	-0.25	0.25

Table 3.3: Upper and lower boundaries of timing parameters in the case where  $n = 5$ .

Parameter	Lower Boundary	Upper Boundary
Epoch1(MJD)	56031.484167333	56045.4437822243
$\Delta\nu_1(\text{s}^{-1})$	$-1.9 \times 10^{-7}$	$1.5 \times 10^{-08}$
$\Delta\dot{\nu}_1(\text{s}^{-2})$	$-5.3 \times 10^{-14}$	$2.7 \times 10^{-14}$
$\Delta\phi_{g,1}$	-0.5	0.5
Epoch2(MJD)	56060	56145
$\Delta\nu_2(\text{s}^{-1})$	0	$1.48 \times 10^{-07}$
$\Delta\dot{\nu}_2(\text{s}^{-2})$	$-2.9 \times 10^{-14}$	$5.1 \times 10^{-14}$
$\Delta\phi_{g,2}$	-0.5	0.5

Table 3.4: Upper and lower boundaries of timing parameters in the case where  $n = 10$ .

### 3.2.2 Comparing the Models

To calculate the evidence, a nested sampling code was applied with a stopping criterion set to equal  $e^{-5}$  – i.e. when new live points made an additional contribution to the evidence that was smaller than a fraction  $e^{-5}$  of the total, the nested sampling code was stopped. The value of  $n$  used for setting our priors was initially taken to be 10 – i.e. far beyond the  $5\sigma$  region. The two models were found to have evidence values of  $\sim e^{-33}$  and  $\sim e^{-29}$  respectively, which yields a Bayes Factor of  $42.5 \pm 3.4$  in favour of the successive anti-glitch model over the anti/normal glitch pair model. The uncertainty on the Bayes Factor was calculated using equation 2.34. According to the definition of [163], a Bayes Factor larger than 40 is already very strong evidence in favour of the successive anti-glitch model (*c.f.* table 2.1).

For each model the posterior distribution of the model parameters was resampled appropriately from our nested sampling results. In figure 3.2 we show posteriors for the parameters of the second event in the double anti-glitch model. The contour lines correspond to 68.3%, 95.5% and 99.7% credible intervals. The maximum posterior corresponds to the following best-fitting parameter values: epoch = MJD 56088.4;  $\Delta\nu = -8.2 \times 10^{-8}$  Hz;  $\Delta\dot{\nu} = 5.2 \times 10^{-15}$  Hz/s; phase change =  $-0.012$  cycles.

### 3.2.3 Robustness Check

We tested the robustness of our results by changing the width of our uniform priors and re-running the nested sampling analysis. Table 3.5 shows the mean Bayes Factor obtained as  $n$  is changed from 10 to 5 to 3. We see that the Bayes Factor fluctuates around a value of  $\sim 45$ , but in all cases our conclusions are consistent. Such a value is also consistent with the estimation from multiple runs as illustrated in table 3.6. The uncertainty values estimated between table 3.5 and table 3.6 are slight different, but at most they differ by a factor of two. The interpretation of the value is displayed in table 2.1 [163]. A Bayes Factor value of  $\sim 45$ , as it's located in the region of  $(10^{3/2}, 100)$ , demonstrates a “very strong support” of the first model, here being the successive anti-glitches model.

The parameter space has been changed to check the influence on the conclusion. Since the  $\Delta\phi_g$  for both timing events in both models were not estimated in [162], we set these phase change parameter to zero and the Bayes Factor increases to around 375, which shows an even stronger support for the successive anti-glitches model.

The timing residuals, after subtracting the best fitting double anti-glitch model, are shown in the upper panel of figure 3.3. This consistency confirms that two anti-glitch events can explain the observed data well.

So far, there have been some physical models proposed in order to explain the putative anti-glitch event (see 3.3 for detailed discussion). Some authors (e.g. [207, 208]) have suggested

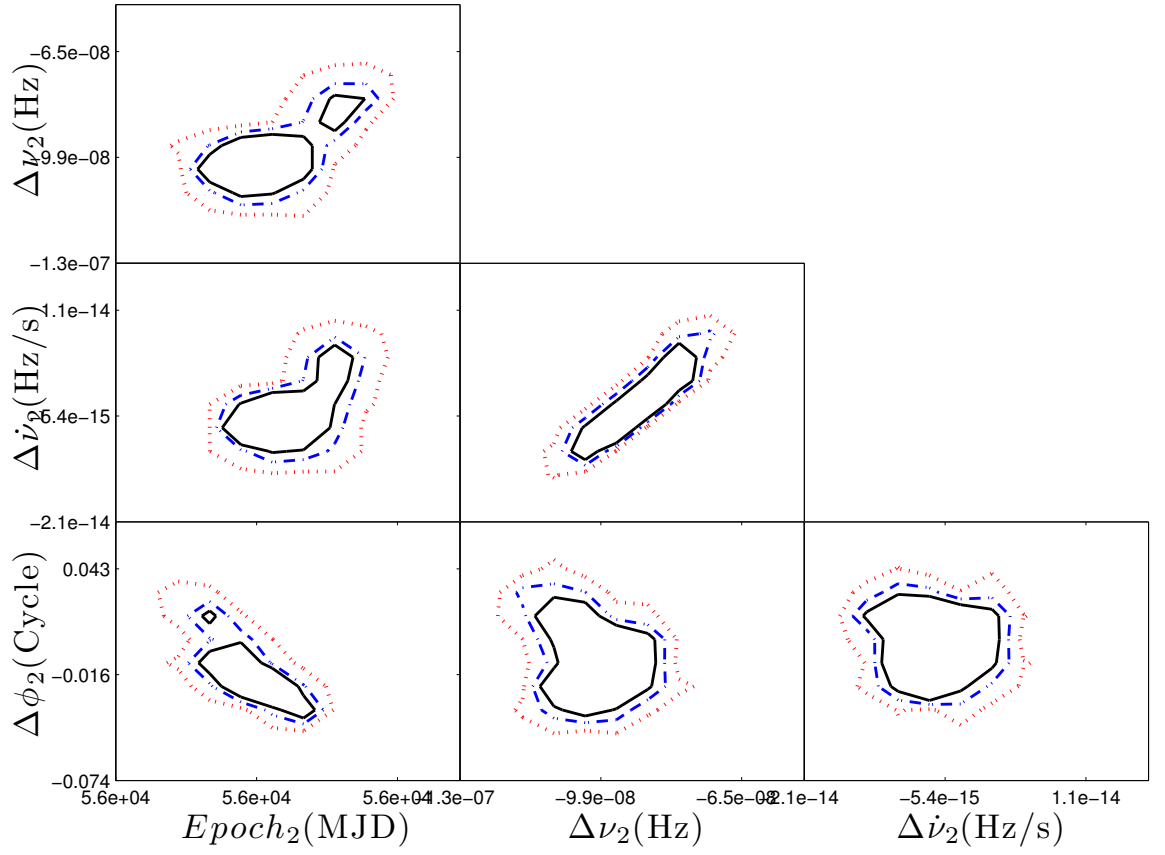


Figure 3.2: Parameter posterior contours for the second anti-glitch event with the successive anti-glitch model, showing 68.3% (solid), 95.5% (dash-dotted) and 99.7% (dotted) credible intervals, based on  $\sim 500$  points resampled from the nested sampling samples.



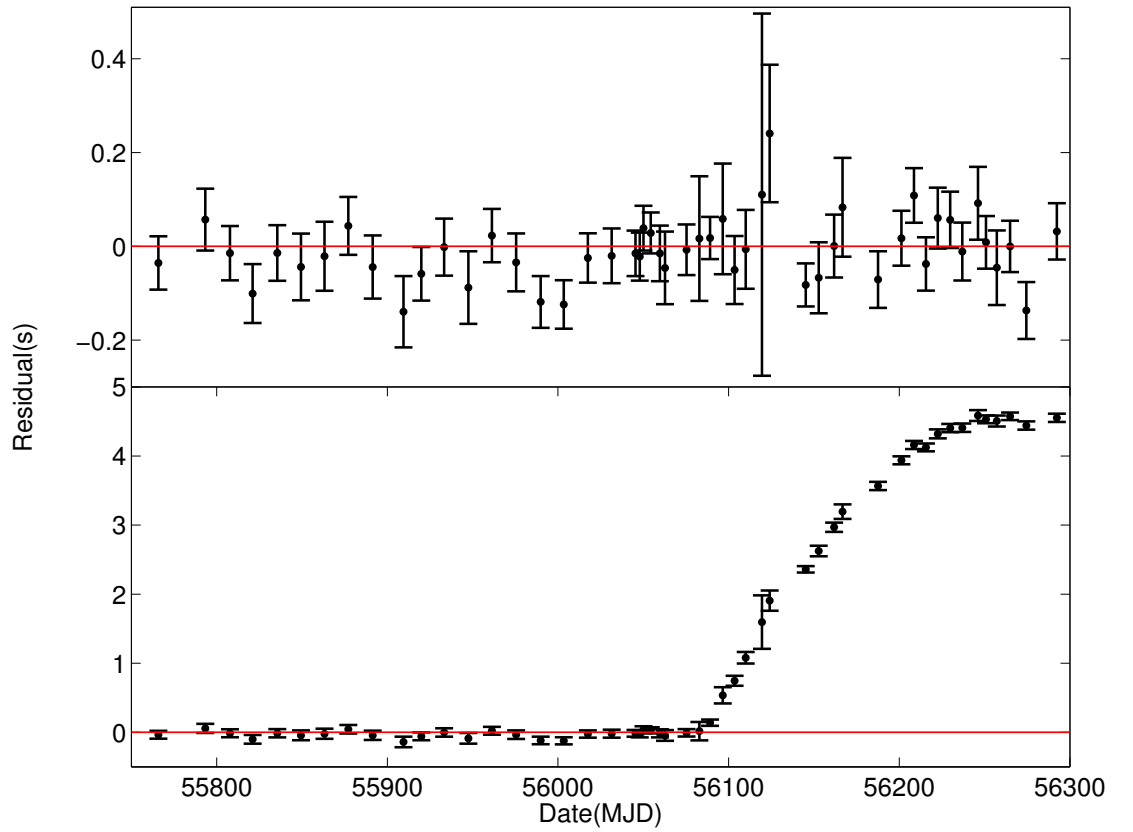


Figure 3.3: Upper panel: Timing residuals of the observed data for two successive anti-glitch events, with best fit parameters as determined by our analysis. Lower panel: Timing residuals of the observed data for only one anti-glitch event. Clearly, two anti-glitch events better explain the data than having a single anti-glitch.

Table 3.5: Estimated Bayes Factors and evidences with uncertainties, obtained from different prior ranges on the model parameters, represented by different  $n$  values (for the definition of  $n$  see main text). A Bayes Factor of around 45 is obtained in each case, indicating consistently strong evidence favouring a successive anti-glitch scenario over an anti/normal glitch pair. The natural logarithm of evidence were consistently estimated with an uncertainty of around 0.05.

$n$ value	10	5	3
Bayes Factor	$42.4 \pm 3.4$	$43.5 \pm 3.1$	$48.6 \pm 3.4$
Successive Anti-glitch Evidence	$\exp(-28.53)$	$\exp(-27.77)$	$\exp(-27.58)$
Anti-/normal glitch Evidence	$\exp(-32.28)$	$\exp(-31.54)$	$\exp(-31.46)$

that the second timing event is not consistent with these physical models, and moreover have questioned whether the observational evidence for the second event is strong enough in the first place. However, as shown in the lower panel of figure 3.3, if we consider only the first event the timing residual will quickly diverge away from zero thereafter, thus showing strong support for the existence of a second timing event. Note that the timing residuals for an anti/normal glitch pair model are also similar to the upper panel in figure 3.3, further supporting the case for a second timing event (but emphasising that to distinguish between an anti/normal glitch pair and an anti-glitch pair is less straightforward).

We also applied model selection to the case of two timing events versus 1 anti-glitch, and the Bayes Factor was found to be  $e^{208}$  – i.e. overwhelmingly favouring the 2 events scenario. This result demonstrates how the Bayes Factor can favour a more complicated model, notwithstanding that it may require additional parameters, when the data are of sufficient quality and a simpler model cannot give a satisfactory fit.

Finally a batch of simulated glitch-free residual data was also generated, with each point drawn from a Gaussian distribution using means and standard deviations from TEMPO2 fits and TOA errors. Nested sampling was applied to this simulated data, and the Bayes factor was computed for the comparison of a successive anti-glitch model and an anti/normal glitch pair model. We calculated 15 Bayes Factor ratios based on 15 realisations of fake anti/normal glitch-free data. We found that the Bayes Factors fluctuated around unity, i.e.  $\exp(0.17 \pm 0.33)$ , showing that intrinsic randomness in glitch-free data will not cause a preference of one model over the other.

Notice that in table 3.5 the uncertainty is estimated from the Nested Sampling algorithm, and might be not representative. We therefore carried out different nested sampling runs with different random seeds, and list the corresponding results in table 3.6. Notice that although the uncertainties are slightly different between estimation from nested sampling and from standard deviation calculation, the mean values are consistent, and the values are generally independent of the choice of  $n$  value.

Table 3.6: Bayes Factors of different realisations with mean and standard deviation. All realisations demonstrate a strong support for the successive anti-glitch events model, the estimated Bayes Factor fluctuate around 45 with standard deviation  $\sim 4$ . The standard deviation is slightly different from the estimation from equation 2.34, while the mean value is consistent. The standard deviations of both models' evidence are also provided.

$n$ value	10	5	3
Realisation 1	41.3	43.5	46.1
Realisation 2	42.4	43.5	46.1
Realisation 3	53.0	44.4	50.1
Realisation 4	42.5	47.9	53.3
Realisation 5	41.1	41.1	51.8
Realisation 6	38.2	40.5	49.2
Realisation 7	45.3	42.6	49.2
Realisation 8	51.5	44.5	47.3
Realisation 9	38.6	39.0	47.8
Realisation 10	46.4	39.5	49.5
Realisation 11	54.4	41.6	49.5
Realisation 12	46.2	43.6	46.8
Realisation 13	44.2	42.3	45.9
mean	45.0	42.6	48.7
standard deviation	5.2	2.4	2.3
successive anti-glitch evidence stdev	0.04	0.05	0.04
anti-/normal glitch evidence stdev	0.08	0.03	0.02

### 3.3 Discussion and Conclusions

We have shown that a model with two successive anti-glitches better explains the observed pulsar data presented in [162] when compared with an anti/normal glitch pair model. Our analysis was robust against variations in the prior ranges, with a Bayes Factor consistently larger than 40 in favour of two anti-glitches. Meanwhile, the Bayes Factor between two events and one event is very large ( $e^{208}$ ), showing conclusively that the two events scenario is favoured over one event.

Prior to the discovery of an anti-glitch there were already several published papers presenting mechanisms that could cause enhanced spin-down, while after its discovery a number of further mechanisms have been proposed seeking to explain its physical origin. Roughly speaking, we can divide the proposed mechanisms into three groups: *internal*, *accretion* and *magnetospherical*.

The internal mechanism is related to that causing glitches in normal pulsars, which can often be satisfactorily explained by the coupling of the crust with the inner faster-rotating superfluid [216], where for a normal pulsar the superfluid interior could not spin slower than the crust. However, as the observed object 1E 2259+586 is a magnetar, where the dominant source of free energy is magnetism instead of rotation, the spin evolution could be vastly

different from that of normal pulsars. [217] suggested that a magnetar could drive differential rotation, which allows a lag in the rotation of the superfluid interior. A sudden rearrangement of the inner structure could induce the interior and crust to corotate again, which would be observed as a sudden spin-down, or anti-glitch [162]. Another possible explanation for the faster-rotating crust might be the twist of a crust patch. As the superfluid vortex is pinned to the crust, a plastic deformation for such a patch will lead to a slower rotating superfluid. A rapid twist would correspond to a conventional spin-up glitch, similar to a normal pulsar counter-part. However, while a gradual twist would have little effect on the secular spin evolution, a rapid unpinning of the associated vortices would give a sudden spin-down, or anti-glitch [218, 205]

Accretion mechanisms suggest that the anti-glitch is caused by the accretion of retrograde material from either a Keplerian ring [211] or from an asteroid [210]. Besides retrograde accretion [209], also proposes an enhanced propeller effect to explain the anti-glitch. Although most accretion models are able to explain both events during the observation, either with or without being accompanied by radiation, this mechanism does not fit the model of magnetars, which has already been supported by many observations. (e.g. [215]).

Magnetospheric models (e.g. [207, 208, 205]) explain the observed anti-glitch with either an enhanced particle wind or a twisting of the magnetic field lines. Although these models fit the observational data for magnetars, most magnetospheric explanations are accompanied with strong radiation and/or a change in pulse profile – neither of which were observed during the second timing event for 1E 2259+586. The magnetospheric mechanism is not favoured since figure 3.3 shows that our analysis strongly favours the existence of the second event.

Among these three mechanisms, our analysis shows that the internal mechanism is most favoured. We note that a satisfactory model should be able to explain the two successive anti-glitches that happened within a relatively short period. If the sudden unpinning of the quantum vortex due to the twist of crust patch is responsible for the anti-glitch, for example, then the gradual plastic deformation of the crust patch should be able to accumulate enough angular momentum within a timescale of several months. If the two anti-glitches are caused by the same mechanism, then the observations may put some constraints on that mechanism. Enhanced radiation, pulse profile changes and enhanced spin-down were observed for the first event while none of these phenomena was observed for the second. Future observations of similar phenomena with higher timing accuracy and sampling frequency will be helpful in order to more fully understand the mechanism responsible.

In this work we've demonstrated how applying Bayesian Inference methods to astrophysical problems can help to distinguish between multiple competing models. The prior is made fair to all models so that the conclusion is not influenced by subjective choice of prior.<sup>1</sup>

---

<sup>1</sup>This work is published in the ApJ. Let., and I am the first author as well as the corresponding author.

## Chapter 4

### Mixed MCMC

One of the main targets of Bayesian inference is to estimate the posterior distribution of desired parameters. To estimate posterior distributions, one naive solution is to use an exhaustive algorithm to calculate the posterior over a dense grid of points in the parameter space. Such brute-force methods will have little or no practical value when dealing with medium-to-high dimensional problems since the computational burden will be prohibitively high. For such problems the ability to concentrate sampling in regions where the posterior probability is high is very important if we are to implement Bayesian inference methods efficiently.

Recall from chapter 2.4, we’ve discussed methods such as *Markov Chain Monte Carlo (MCMC)* and *Nested Sampling* and the several variants. These methods are well tailored to explore the posterior distribution over high dimensional parameter spaces. While the computational cost of brute-force methods increases exponentially with the dimension, MCMC usually only grows slowly with dimension [176, 203].

Generally, the method of MCMC works well so long as the posterior surface is sufficiently smooth. However, when the posterior distribution has a complicated structure, MCMC will become inefficient. For example, MCMC samplers are known to get “caught” in a local mode of the posterior, and unable to jump out and explore any other isolated modes in the parameter space [219, 193]. So a lot of methods have been proposed to make the MCMC sampling more efficient (*e.g.* [180, 198, 196])

In this chapter, we discuss a novel method, *mixed MCMC*, to deal with such issues. The conventional MCMC algorithm is robust for exploring the detailed structure of the posterior surface, and we want to retain that property while enabling some global “communication” between different regions of the parameter space so that the sampler can make jumps between those regions without requiring a very long exploration time.

Readers are reminded that we design the algorithm of mixed MCMC so that it’s tailored for

multi-modal posteriors with some limited knowledge of the posterior. As discussed in chapter 4.2, the implementation of mixed MCMC would require some rough information about the posterior, which naturally limit the application of the method. However, time-consuming algorithms could be used to assist obtaining such information, and there are situations where these information is already available.

## 4.1 Method

As we recall from chapter 2, the algorithm of MCMC sets out to sample a chain of points in the parameter space and at the  $i^{th}$  iteration (i.e. after  $i - 1$  points have already been sampled) a *candidate* point  $\theta^*$  is randomly sampled from some specified proposal distribution, based solely on the position of the *previous point* in the chain  $\theta^{(i-1)}$ . The corresponding posterior for this candidate point is calculated, and compared with the posterior at  $\theta^{(i-1)}$ . The candidate is accepted only with a certain *acceptance probability* (see chapter 2.4). One can observe that the sampling will generally proceed “uphill” – i.e. to regions of the parameter space where the value of the posterior is larger – while sometimes it can also go “downhill” to regions where the posterior takes on lower values [156, 157, 177, 154].

### 4.1.1 Markov Chain Monte Carlo

Hereafter, we define the posterior  $f(\theta) = p(\theta|D, I)$ , the prior  $\pi(\theta) = p(\theta, I)$  and likelihood  $\ell(\theta) = p(D|\theta, I)$ , where  $\theta$  is the parameter set,  $D$  is the data and  $I$  is the information.

The simplest form of Markov Chain Monte Carlo (MCMC) is known as the Metropolis algorithm, which can be achieved by the following steps [156, 157, 177, 220].

1. Arbitrarily choose a starting point  $\theta^{(0)}$  that satisfies  $f(\theta^{(0)}) > 0$ , and a symmetric proposal distribution  $J(\theta_a|\theta_b)$ . Set step index  $i=0$ .
2. Increment  $i$  by 1.
3. Randomly propose a new parameter set  $\theta^*$  by sampling from  $J(\cdot|\theta^{(i-1)})$ .
4. Calculate the *Metropolis ratio* given by

$$r = \frac{f(\theta^*)}{f(\theta^{(i-1)})} \quad (4.1)$$

5. Accept the proposed parameter set  $\theta^*$  with *acceptance probability*

$$\alpha(\theta^{(i-1)}, \theta^*) \triangleq \min(1, r) \quad (4.2)$$

If  $r \geq 1$ , then the candidate is accepted, so the new point is  $\theta^{(i)} = \theta^*$ .

If  $r < 1$ , draw a random number  $rand$  from a uniform distribution  $U[0, 1]$ , and if  $rand < r$ , then set  $\theta^{(i)} = \theta^*$ ; otherwise set  $\theta^{(i)} = \theta^{(i-1)}$ .

Steps 2-5 are repeated until a large enough number of points have been sampled. This termination could be controlled by a preset number, or by monitoring the samples' distribution and check if it's sufficiently stable [185]. The beginning period, which is generally called as the “burn-in” stage, is discarded to prevent the influence of the arbitrary choice of starting point  $\theta^{(0)}$ .

The Metropolis-Hastings (M-H) algorithm is a more general form of the Metropolis algorithm. In the Metropolis algorithm, the proposal distribution is symmetric, that is  $J(\theta_a|\theta_b) = J(\theta_b|\theta_a)$ , but this condition is not necessary. In the M-H algorithm we relax this symmetric condition, so that equation (4.1) should be modified as follows

$$r = \frac{f(\theta^*)J(\theta^{(i-1)}|\theta^*)}{f(\theta^{(i-1)})J(\theta^*|\theta^{(i-1)})}. \quad (4.3)$$

It is clear that when the proposal distribution is symmetric, equation (4.3) is identical to equation (4.1).

It can be shown that the number density of the sampled points will represent a sample from the posterior distribution [156]. Thus estimation of the parameter(s) that characterise the posterior distribution becomes possible with a sufficiently large number of sampling points.

## 4.2 mixed MCMC

If the starting point and/or the proposal density is not properly chosen, the MCMC sampler might become stuck in a local mode, and will not be able to appropriately explore the whole parameter space. This might introduce a statistical bias in the parameter estimation carried out by MCMC, particularly when the target distribution is multi-modal. This motivates the realisation of mixed MCMC as a really Markovian realisation of MCMC that can sample the posterior efficiently [193, 185, 202].

Here we propose a novel method which we term *mixed MCMC* to perform Bayesian inference on multi-modal posterior distributions. This method can allow the sampler to communicate between different local maxima, so that the sampler will be able to represent local peaks, as well as to explore the global structure. As noted previously, our method requires some limited information about the location of the multiple modes before sampling. In many cases, however, we will have at least some rough prior knowledge about the posterior, and

we can use this information to guide the sampler. Even in the absence of such prior knowledge, other existing global sampling methods can be tailored for this purpose to speed up this process [193, 181, 221].

### 4.2.1 Algorithm

The main difference between the algorithm for mixed MCMC and the conventional MCMC algorithms simply roots in the use of a novel form of proposal density. The sampler should be able to generate candidates from different sub-regions, while a proper choice of Metropolis ratio will ensure that the sampling between those different sub-regions satisfies detailed balance.

Suppose, as a result of existing prior knowledge, or with the help of some other global sampling method, we have some information about the posterior distribution that is sufficient to identify the existence and the rough location of the several modes in posterior distribution, where the location of the  $t^{th}$  mode is labelled as  $\theta_t^0$ . We can then divide the parameter space into several distinct *sub-regions* each of which we assume contains a single mode of the posterior [156, 200].

We should bear in mind that this method is designed for multi-modal posteriors, thus the proposal density should be designed in a way that it can propose new candidates in all posterior modes. Thus we assign to the  $t^{th}$  sub-region what we term a *picking up probability*,  $p_t$ , which determines the probability to get a new candidate in the  $t^{th}$  sub-region. Ideally, this probability should be proportional to the marginal likelihood (also known as the evidence) within the sub-region – i.e. the probability that the candidate point lies within that sub-region. Note also that the picking up probability should satisfy the normalisation requirement  $\sum_t p_t = 1$ . At the same time it will maximise the efficiency of our approach if  $p_t \propto \int_{V_t} f(\theta) d\theta$ , where  $V_t$  is the volume of the  $t^{th}$  sub-region of the parameter space.

Suppose we decide to generate a candidate point in the  $t^{th}$  sub-region, while the current (i.e. most recently updated) point  $\theta^{(i-1)}$  is located in the  $s^{th}$  sub-region. Then a normalised multivariate distribution (most conveniently taken to be a Gaussian) centering around the point  $\theta_t^0 - \theta_s^0 + \theta^{(i-1)}$  is used as proposal density, and a candidate is drawn from this distribution. After calculating the value of the posterior at this candidate point, and then computing the Metropolis ratio,  $r$ , in the usual way, we can decide to accept the candidate point with the acceptance probability  $\alpha$  as before.

In more detail our mixed MCMC algorithm can be illustrated with the following pseudo-code.

1. Obtain some rough approximation to the posterior distribution using other methods.



Identify  $m$  modes in the parameter space, and estimate their central locations given by  $\theta_t^0$ .

2.  $p_t$  to be the picking up probability, defined as proportional to the volume of the  $t^{th}$  sub-region, with  $\sum_t p_t = 1$ . Set step label  $i = 0$
3. Randomly pick a starting point,  $\theta^{(i)}$ .
4. *while*(not converged)
  - (a) Set  $i=i+1$
  - (b) Randomly pick a sub-region number  $t$  with probability  $p_t$  and assign  $s$  to be the current sub-region index.  $s, t \in \{1, \dots, m\}$  where  $m$  is the number of all sub-regions.
  - (c) Generate the candidate point  $\theta^* = \theta^{(i-1)} + \theta_t^0 - \theta_s^0 + \delta\theta$  drawn from the proposal density  $\sim J(\cdot | \theta^{(i-1)})$
  - (d) Calculate the Metropolis ratio  $r$  based on the candidate and the previous point,
 
$$r = \frac{f(\theta^*)_{p_t}}{f(\theta^{(i-1)})_{p_s}}.$$
  - (e) Generate a random number  $\text{rand} \sim U[0, 1]$ .
  - (f) Accept the proposed parameter set  $\theta^*$  with *acceptance probability*  $\alpha = \min(1, r)$  as follows:
 
$$\begin{array}{ll} \text{if}(r > \text{rand}), & \text{update, } \theta^{(i)} = \theta^* \\ \text{else} & \theta^{(i)} = \theta^{(i-1)} \end{array}$$

The mixed MCMC algorithm set out above is strictly Markovian, and detailed balance is achieved by construction. Thus the number of points sampled in given sub-region should provide an estimate of ratio between local evidences. As noted above, in order to maximise the efficiency of the algorithm the picking up probability  $p_t$  should better be proportional to the local evidence.

Also, we can notice that when the proposed point and the previous point are located in the same sub-region, then the algorithm reduces to the conventional M-H algorithm, which further verifies its validity.

One needs to take care to allow certain burn-in stage before allowing the mixed MCMC sampling, if the  $\theta_t^0$  apart too far away from the peak, the sampler might stop sampling at all in the corresponding mode.

## 4.3 Toy Model

We demonstrate our mixed MCMC algorithm using a simple toy model. On a two dimensional  $x - y$  parameter space, we considered a posterior distribution as the sum of a pair of well-separated bivariate normal distributions. Where the parameters are  $x$  and  $y$ , and the two artificial posterior modes locate in  $(\mu_1^x, \mu_1^y)$  and  $(\mu_2^x, \mu_2^y)$ , each mode can be described by a bivariate normal distribution with a diagonal covariance matrix, where the standard deviation in each direction is  $\sigma_1^x, \sigma_1^y$  for the first mode and  $\sigma_2^x, \sigma_2^y$  for the second mode. The form of this posterior is, therefore:

$$\begin{aligned} f(x, y) = & C_1 \exp\left[-\frac{(x - \mu_1^x)^2}{2(\sigma_1^x)^2} - \frac{(y - \mu_1^y)^2}{2(\sigma_1^y)^2}\right] \\ & + C_2 \exp\left[-\frac{(x - \mu_2^x)^2}{2(\sigma_2^x)^2} - \frac{(y - \mu_2^y)^2}{2(\sigma_2^y)^2}\right]. \end{aligned} \quad (4.4)$$

The coefficients  $C_1$  and  $C_2$  allow the two modes to differ in height, and when integrated over the entire parameter space the normalisation condition implies that

$$\int_{x,y} dx dy f(x, y) = 1. \quad (4.5)$$

For simplicity, we chose  $\mu_1^x = -\mu_2^x = -3$ ,  $\sigma_1^x = \sigma_2^x = 0.1$ ,  $\mu_1^y = \mu_2^y = 0$ ,  $\sigma_1^y = \sigma_2^y = 0.1$  and the ratio of two coefficients  $C_1 : C_2$  is kept as 1 : 3.

In this toy model test, we only concentrate on the validity of the mixed MCMC method, and do not consider in detail other factors such as its efficiency or generality. Thus, we assume prior knowledge of the separated structure of the posterior distribution. Given this assumption, it is possible to analytically calculate the  $\Delta\chi^2$  value that corresponds to the contour within which a certain fraction of the entire volume of the posterior is located, thus providing us with an exact theoretical reference result with which to compare. For a one-dimensional Gaussian distribution, the  $1\sigma$ ,  $2\sigma$  and  $3\sigma$  credible regions correspond to 68.27%, 95.45% and 99.73% of the cumulative probability function (CDF) respectively. Thus, it is convenient to consider for our toy model posterior the  $\Delta\chi^2$  values that correspond to the 68.27%, 95.45% and 99.73% of the CDF, and compare it with the sample estimates obtained from application of our mixed MCMC algorithm.

Under the Gaussian assumption,  $\Delta\chi^2 = -2\log(\mathcal{L})$ , so in the general case we will compare the value of  $2\Delta\log(\mathcal{L})$  with its theoretical evaluation. The posterior of the toy model can be taken as two independent bivariate Gaussians, each with a diagonal covariance matrix.

### 4.3.1 Analytical Evaluation of $2\Delta \log(\mathcal{L})$

In this section we present the calculation of  $2\Delta \log(\mathcal{L})$ , which is defined as

$$2\Delta \log(\mathcal{L}) = 2\log(\mathcal{L}_{max}) - 2\log(\mathcal{L}),$$

for our toy model posterior.

In the toy model, there are two well separated modes. The two modes have an evidence ratio of  $w_1 : w_2$ , where  $w_1 + w_2 = 1$  is the normalisation requirement.

We define the two independent parts of the posterior as

$$P_1(x, y) = \frac{1}{2\pi\sigma_1^x\sigma_1^y} \exp\left[-\frac{(x - \mu_1^x)^2}{2(\sigma_1^x)^2} - \frac{(y - \mu_1^y)^2}{2(\sigma_1^y)^2}\right]$$

$$P_2(x, y) = \frac{1}{2\pi\sigma_2^x\sigma_2^y} \exp\left[-\frac{(x - \mu_2^x)^2}{2(\sigma_2^x)^2} - \frac{(y - \mu_2^y)^2}{2(\sigma_2^y)^2}\right]$$

and the posterior can be written as

$$f(x, y) = w_1 P_1(x, y) + w_2 P_2(x, y) \quad (4.6)$$

The peak values of the posterior for its two modes are  $C_1 \triangleq \frac{w_1}{2\pi\sigma_1^x\sigma_1^y}$  and  $C_2 \triangleq \frac{w_2}{2\pi\sigma_2^x\sigma_2^y}$  respectively.

For simplicity, we replace  $\frac{(x - \mu_1^x)^2}{(\sigma_1^x)^2} + \frac{(y - \mu_1^y)^2}{(\sigma_1^y)^2} = r_1^2$  and  $\frac{(x - \mu_2^x)^2}{(\sigma_2^x)^2} + \frac{(y - \mu_2^y)^2}{(\sigma_2^y)^2} = r_2^2$  and rewrite the posterior as

$$f(x, y) = C_1 \exp(-r_1^2/2) + C_2 \exp(-r_2^2/2) \quad (4.7)$$

Without losing generality, we assume  $C_1 > C_2$ , and so the highest posterior value  $f_{max} = C_1$ , and highest posterior value for the secondary peak is  $C_2$ . We define  $r_0$  as  $C_1 \exp(-r_0^2/2) = C_2$ , equivalently,  $\exp(-r_0^2/2) = \frac{C_2}{C_1}$ .

Our aim is to find the expression for  $\Delta\chi^2(C)$ , so that given  $C$ , we have

$$\int_{f > \exp(-\frac{\Delta\chi^2}{2})} f(x, y) \, dx \, dy = C$$

When  $f > \exp(-r_0^2/2)$ ,

$$\begin{aligned}
C &= \int_{f > \exp(-\Delta\chi^2/2)} C_1 \exp(-\frac{r^2}{2}) dx dy \\
&= 2\pi C_1 \sigma_1^x \sigma_1^y \int_0^{\Delta\chi^2} \exp(-\frac{r^2}{2}) dr^2 / 2 \\
&= w_1 [1 - \exp(-\Delta\chi^2/2)].
\end{aligned} \tag{4.8}$$

This expression is valid so long as  $C < C_0 \triangleq w_1(1 - \frac{C_2}{C_1}) = w_1 - w_2 \frac{\sigma_1^x \sigma_1^y}{\sigma_2^x \sigma_2^y}$ ; if, however,  $C$  is bigger, than we have to include the contribution from the secondary mode.

$$\begin{aligned}
C &= C_0 + \int_{r_0}^{r_1} w_1 \exp(-\frac{r^2}{2}) r dr + \int_0^{r_2} w_2 \exp(-\frac{r^2}{2}) r dr \\
C - C_0 &= w_1 [\frac{C_2}{C_1} - \exp(-r_1^2/2)] + w_2 [1 - \exp(-r_2^2/2)] \\
C - w_1 + w_2 \frac{\sigma_1^x \sigma_1^y}{\sigma_2^x \sigma_2^y} &= w_2 \frac{\sigma_1^x \sigma_1^y}{\sigma_2^x \sigma_2^y} - w_1 \exp(-r_1^2/2) \\
&\quad + w_2 - w_1 \frac{\sigma_2^x \sigma_2^y}{\sigma_1^x \sigma_1^y} \exp(-r_1^2/2) \\
C &= 1 - w_1 (1 + \frac{\sigma_2^x \sigma_2^y}{\sigma_1^x \sigma_1^y}) \exp(-r_1^2/2)
\end{aligned} \tag{4.9}$$

In the third line we used the relation that  $w_1 P(r_1) = w_2 P(r_2)$ .

Furthermore, we have

$$\begin{aligned}
\exp(-\frac{r_1^2}{2}) &= \frac{1 - C}{w_1 (1 + \frac{\sigma_2^x \sigma_2^y}{\sigma_1^x \sigma_1^y})} \\
-r_1^2/2 &= \log(1 - C) - \log(w_1) - \log(1 + \frac{\sigma_2^x \sigma_2^y}{\sigma_1^x \sigma_1^y}) \\
\Delta\chi^2 = r_1^2 &= -2[\log(1 - C) - \log(w_1) \\
&\quad - \log(1 + \frac{\sigma_2^x \sigma_2^y}{\sigma_1^x \sigma_1^y})]
\end{aligned} \tag{4.10}$$

We determined  $w_1 = \frac{3}{4}$  and  $w_2 = \frac{1}{4}$ , while keeping  $\sigma_1^x = \sigma_2^x$  and  $\sigma_1^y = \sigma_2^y$ , and choosing the  $C$  value as 68.27%, 95.45% and 99.73%. This yields the corresponding  $\Delta\chi^2$  values as 3.11, 6.99 and 12.64.

Table 4.1: Likelihoods corresponding to 68.27%, 95.45% and 99.73% credible regions and the portion of points from the primary mode.  $\sigma_1^x = \sigma_1^y = \sigma_2^x = \sigma_2^y = 0.1$ .

credible region	likelihood			portion
	68.27%	95.45%	99.73%	
Realisation 1	3.19	6.87	11.30	74.2%
Realisation 2	3.02	6.99	13.01	74.6%
Realisation 3	3.15	7.00	12.79	74.8%
Realisation 4	3.19	6.84	12.40	74.2%
Realisation 5	3.00	6.89	12.43	77.3%
Realisation 6	3.08	7.14	13.32	75.6%
Realisation 7	3.21	7.30	13.77	73.2%
Realisation 8	3.28	7.29	13.96	73.2%
Realisation 9	3.01	6.98	12.63	77.0%
Realisation 10	3.03	7.02	12.22	74.7%
Realisation 11	3.12	6.99	12.63	76.2%
Realisation 12	3.16	7.20	13.11	74.4%
Realisation 13	3.13	6.89	12.60	72.0%
Realisation 14	3.05	6.95	11.97	74.4%
Realisation 15	3.18	6.99	12.78	74.5%
mean	3.12	7.02	12.72	74.7%
standard deviation	0.08	0.14	0.67	1.4%
theoretical	3.11	6.99	12.64	75.0%

### 4.3.2 Application of the mixed MCMC

We generated a chain with  $10^5$  points. The proposal density was set to be a bivariate Gaussian distribution in addition to the shift between sub-regions, with covariance matrix equal to the identity matrix multiplied by  $\sigma = 0.1$ .

For this particular toy model, theoretically the corresponding 68.27%, 95.45% and 99.73% credible regions should have values of  $2\Delta \log(\mathcal{L})$  equal to 3.11, 6.99 and 12.64. A typical realisation gives result as 3.14, 7.04 and 12.24, and the numbers of points sampled in the two sub-regions are 2559 and 7441, which is consistent with the 1 : 3 ratio assumed for the coefficients  $C_1$  and  $C_2$ . In table 4.1 we listed the  $2\Delta \log(\mathcal{L})$  values for 15 different realisations, the mean value of these realisations shows good consistency with the theoretical prediction, and the standard deviation is relatively small. We can also notice that the standard deviation tends to increase with  $2\Delta \log(\mathcal{L})$ , which is expected as this corresponds to larger credible regions, and it subject to larger random fluctuation.

If we modify the standard deviation of one parameter to 0.2 in the primary mode, the theoretical  $2\Delta \log(\mathcal{L})$  corresponds to 68.27%, 95.45% and 99.73% credible regions would change to 2.53, 6.42 and 12.06.  $2\Delta \log(\mathcal{L})$  values for 15 realisations of such modified toy model were presented in table 4.2, again, the mean value shows good consistency, with satisfactory standard deviation which increase with  $2\Delta \log(\mathcal{L})$ .

Table 4.2: Likelihoods corresponding to 68.27%, 95.45% and 99.73% credible regions and the portion of points from the primary mode.  $\sigma_1^x = \sigma_2^x = \sigma_2^y = 0.1$ ,  $\sigma_1^y = 0.2$ .

credible region	likelihood			portion
	68.27%	95.45%	99.73%	
Realisation 1	2.54	6.30	12.06	76.9%
Realisation 2	2.63	6.52	12.22	72.8%
Realisation 3	2.51	6.50	11.24	76.8%
Realisation 4	2.59	6.57	11.58	75.2%
Realisation 5	2.50	6.18	11.13	76.6%
Realisation 6	2.56	6.32	11.49	74.8%
Realisation 7	2.52	6.11	12.44	72.0%
Realisation 8	2.50	6.03	11.42	75.8%
Realisation 9	2.44	6.41	11.74	74.2%
Realisation 10	2.48	6.37	11.42	74.8%
Realisation 11	2.52	6.45	11.85	74.7%
Realisation 12	2.58	6.55	12.03	74.5%
Realisation 13	2.55	6.45	12.18	73.7%
Realisation 14	2.55	6.42	12.42	73.2%
Realisation 15	2.44	6.42	12.08	75.4%
mean	2.53	6.37	11.88	74.8%
standard deviation	0.05	0.16	0.47	1.4%
theoretical	2.53	6.42	12.06	75.0%

In figure 4.1 an example of the sampling results is shown, with blue, green and red colour points representing the highest (i.e. largest value of the posterior) 68.27%, 95.45% and 99.73% fraction of the samples, after sorting the posterior values in descending order.

## 4.4 Discussion

A novel method, which we call *mixed MCMC*, has been proposed. In the situations when the multi-modal characteristics of the posterior distribution are already roughly known, the parameter space can be split into several sub-regions, each of which hosts a single mode, and our mixed MCMC method can be applied. The proposal density can generate candidates in different sub-regions by adding a shift from the current sub-region to the proposed new sub-region. In this way, a comparison between different sub-regions can be done globally, which improves efficiency. This algorithm is strictly Markovian, so the detailed balance requirement is fulfilled. The concept of mixed MCMC is realised by enabling proposed candidate points to be generated from different modes of the posterior. Admittedly, the mixed MCMC approach must rely on other methods to first identify the multiple modes of the posterior distribution. However, since that identification will generally require only rough information, we can expect this initial stage to be rapid. Moreover, other existing methods

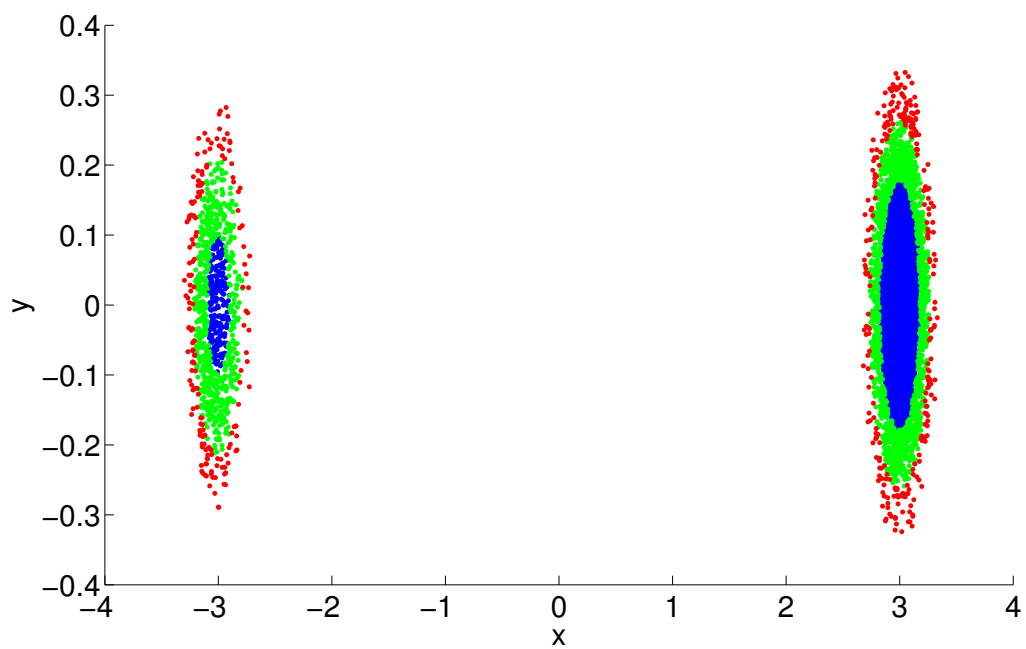


Figure 4.1: Example realisation of mixed MCMC applied to the toy model posterior distribution described in the text. Here blue, green and red colour points represent subsets of the sampled points with the highest 68.27%, 95.45% and 99.73% fractions respectively of the posterior value among all samples. The different areas of the two different modes reflect the fact that the two modes have different weights. Like all MCMC results, the density of the sampling points reflect the posterior.

already provide some solutions to the problem of identifying multiple modes [168, 181].

From another perspective, if we view separate parameter subspaces as different models, this mixed MCMC algorithm can be viewed as an special form of reversible jump MCMC [160, 197], which can sample from different models even when they have different dimensionality, and thus provides the Bayesian odds ratio of two models.

This method is a novel realisation of MCMC which can achieve high efficiency in analysing multi-modal posterior distributions by virtue of its unique form of proposal density. It relies not only on local information, but also on the global structure through swapping between different modes. In particular the candidate point is accepted with an *acceptance probability*  $a = \min(1, r)$ , where  $r$  is the *Metropolis ratio*, which takes into account the global information about the multiple modes of the posterior.

So far, we have not discussed in detail how to obtain rough information about the posterior modes. We note that methods such as MultiNest [168, 181] aim to solve similar problems, so their approach could be directly applied here. Some other methods like parallel tempering MCMC [193] or *k-means* [222] can also be modified and applied here.

We leave the detailed comparison with other methods, like parallel tempering MCMC to future work. However, by not throwing away points in parallel chains, and the design of the proposal density to have a relatively short autocorrelation length we expect the mixed MCMC algorithm to be quite efficient [199].

We also notice that the efficiency of simple random walk MCMC scales  $\mathcal{O}(d^2)$  with the dimensionality  $d$ , while for the Hamiltonian MCMC methods the efficiency scales with  $\mathcal{O}(d^{5/4})$  [223]. We anticipate in future work to modify the algorithm of mixed MCMC to the Hamiltonian MCMC to gain more efficiency in higher dimensional models.

In this work, we also applied our method to a simple toy model, with two distinct well-separated modes, to demonstrate its efficacy. With  $10^5$  samples, our mixed MCMC was able to both find the picking up probability, which represents the bulk distribution of the posterior (i.e. the probability of belonging to each mode) and also the Bayesian credible regions for the posterior as a whole – each of which show excellent agreement with the exact, theoretically computed values for our toy model.

We should notice, however, that in more general cases, the posterior won't necessarily have modes with similar size and shape in the parameter space, which could hinder the application of mixed MCMC. However, for such cases, we can transform the parameter space locally around the maxima, to reshape the multiple modes into similar shapes and sizes.<sup>1</sup>

---

<sup>1</sup>This work was initiated and mainly developed by myself, under the supervision of Prof. Martin Hendry and Dr. Ik Siong Heng.



## Chapter 5

# Global Optimisation for Future Gravitational Wave Detectors Sites

In this chapter, we applied the mixed MCMC methodology (*chapter 4*) to a problem of optimising the sites of future generation gravitational wave detectors. Mixed MCMC is designed to accelerate the sampling speed for a multi-modal posterior, which would rely on some rough information of the locations of multiple modes. As we will explain in details later, the task of optimising GWD sites should view the GWD network as a whole, so that the scientific output of a GWD network can be quantified as a single Figure of Merit. Such a map from multiple parameters (geographical coordinates) to a real number reminds us of the problem of sampling high dimensional posteriors. We notice that the problem is naturally multi-modal, as the Figure of Merit of such a network would remain constant upon a simultaneous shift of the entire network in the same direction. Also geological information could be used to guide the positions of the multiple modes. We thus conclude that mixed MCMC would be an ideal sampling strategy for such a problem.

Recent advances in technology should enable us in the near future to open a new gravitational-wave (GW) window for astronomy. Although no signals have been detected yet, there are excellent prospects for the first detections to take place before the end of the current decade, as the ‘first generation’ GW detectors LIGO [51] and Virgo [52] are upgraded to their ‘second generation’ counterparts Advanced LIGO (aLIGO) [56] and Advanced Virgo (AdV) [57], with an increase of more than a factor of ten in sensitivity, which translates to an increase in detection rate of a thousand [224, 225]. At the same time, detailed design studies for proposed future generation GW detectors such as the Einstein Telescope (ET) have recently been completed [226], and the prospects for multi-messenger astrophysics and cosmology with such instruments have been investigated [227, 2]. It seems clear that the first successful detection of GW signals with aLIGO and AdV will provide tremendous impetus for the nascent field of GW astronomy, and thus generate renewed enthusiasm for the building of

new and even more advanced detectors in the future.

However, the costs of building GW observatories, particularly future generation detectors, are very high [226]. Even in the most optimistic scenarios, therefore, it seems unrealistic to expect that more than (say) half a dozen future generation GW detectors will be built in the coming decades. Consequently, the optimal identification of sites for future GW detectors is an important issue that needs to be carefully considered.

The site selection must take into account many factors such as seismic stability and other sources of gravitational noise [228]. For future generation detectors, in order to achieve a further order-of-magnitude improvement in sensitivity there is a scientific motivation for constructing them underground. This adds considerably to the cost, however, and may involve e.g. the building of extensive tunnels which may in turn place significant constraints on transport infrastructure – all of which will contribute to the overall construction budget. Thus choosing reliable sites for future GW detectors needs to consider a wide range of factors in addition to purely scientific constraints [136].

However, the optimal choice of site will be significantly different when we are planning to optimising multiple detectors instead of a single instrument [137]. Unlike electromagnetic (EM) telescopes, which generally observe only a very small patch of sky at any time, GW detectors have an all-sky response [137, 103]. This property enables multiple GW detectors, when combined as a network, to gain improved information on the source – including its position in the sky [17]. In particular, while it is nearly impossible to localise the sky position of a GW source with only one detector in operation when no additional information (e.g. an EM counterpart) is available, for multiple detectors the difference in arrival time of the GW signal at each detector can be used to localise the source direction. For a given GW event, different configurations of such a network of detectors would provide different information on its sky position [229]. For example, multiple GW detectors that are concentrated in a small geographical area would result in very poor limits on the source’s sky location, while a network of well-separated detectors would provide much tighter constraints. These basic considerations provided a strong original motivation for the decision to build two LIGO detectors in geographically well-separated locations in the US, and more recently have informed the proposal to locate an aLIGO detector in India [60].

Previously *Raffai et al.* [136] investigated constraints on optimal GW detector networks. In that work the optimisation was based on three “Figures of Merit” – discussed in Section 5.1 – and was mainly focused on a 2-detector network. The generalisation to networks with more detectors was carried out by adding one detector at a time, keeping the locations of all previously-sited detectors fixed. After considering  $\sim 1500$  possible sites for the additional detector, covering all allowable regions, the best site was chosen. This method fitted well to the aims of the *Raffai et al.* [136] study, but a simultaneous optimisation of *all* sites in

a  $N > 2$  detector network requires a different approach. Nevertheless, the method already proved to be very useful: for example, considering a five detector network with the first four sites to be aLIGO Livingston, aLIGO Hanford, AdV and KAGRA [61], the optimal location and orientation of a possible fifth advanced detector in India could be determined via the exhaustive exploration of allowable five-detector configurations and the calculation of the appropriate figures of merit for the resulting network corresponding to each configuration. This approach therefore gave indications of the relative merits of different candidate sites and orientations in the planning of aLIGO India [230].

Throughout this chapter, we specify the term *future generation GW detectors* to be ET-like detectors, *i.e.*, with triangular configuration, 10km armlength, expected to be built underground *etc.*, while the term *advanced GW detectors* refers to the detectors like aLIGO or AdV which are being under construction or planned. Our primary interest lies in the optimisation for future generation GW detectors specifically.

In this work, we extend the method of *Raffai et al.* [136] to the case where the optimisation for multiple detectors is carried out simultaneously. This allowed us, for example, to find the optimal sites for a 3-detector or a 5-detector network, treating the locations of *all* detectors in the network as free parameters. Also, we are interested to know if an ideal site for a detector network comprising 3 detectors could still be considered a good location when being part of a network with 5 detectors. Moreover, we want to be able to discuss what the range of possible tolerable configurations might be. Should the optimal configuration not be available because of any unexpected reason, such information would be priceless, and this can only be achieved by a simultaneous optimisation process for the detector sites.

In practice the spatial distribution of GW sources is not isotropic as the nearby galaxies are strongly clustered, and the seismic environment varies greatly with location; however, for simplicity we ignore any such directional dependence here. Hence the comparison of different networks is purely determined by their relative geometric shape, and not by the physical characteristics of the sites' actual geographical locations. Consequently shifting (rotating on Earth) the whole detector network by a small amount while keeping fixed the angles between the individual detectors would not, in our analysis, affect the overall performance of the network. Thus, we acknowledge that to ask in complete generality “what is the optimal network?” is too ill-defined a question. Instead, therefore, we ask a slightly different and more specific question: assuming that we will build 3 (or 5) future generation GW detectors, where are the ideal sites for those detectors that maximise their flexibility – *i.e.* such that they can belong to a large number of different “good” detector networks?

We acknowledge, however, that at least on nearby cosmological scales the expected distribution of sources will be neither isotropic nor homogeneous – although the impact of these effects on a given GW detector network will be mitigated by the rotation of the Earth: a net-

work which might be optimised e.g. to the distribution of galaxies in the Local Supercluster at one time will not in general be optimised a few hours later. Similarly, seismic stability will also depend in detail on geographical location; this will make some network configurations in practice more favourable than others. However, in the following analysis, we ignore such effects for the case of simplicity. We should clarify here that we do not take explicitly into consideration the current or planned future sites of ground based detectors. Although these advanced detectors are expected themselves to make ground-breaking discoveries, their sensitivities are still expected to be an order of magnitude or more lower than that of proposed future generation GW detectors. Thus, in the context of the future generation GW detector network that we are investigating in this study, such advanced detectors would make only a limited contribution. Also, most of the current advanced detector sites are not considered ideal according to our exclusion criteria, as discussed in chapter 5.2.4. As the future generation detectors are expected to be constructed underground, there is no compelling motivation to locate them at the sites of the current detectors. Hence in this study we simply ignore the locations of current or planned advanced detectors.

## 5.1 Figures of Merit for GW Detector Networks

In *Raffai et al* [136] the authors comprehensively investigated different GW detector network configurations, characterising their relative performance using some reasonable Figures of Merit (FoMs). So that we can conveniently compare our results with that previous work, we mostly adopt the same definitions as [136] and consistently construct our FoMs as

- $I$ , which measures the network's capability of reconstructing the source polarisation,

$$I = \left( \frac{1}{4\pi} \int |F_+^{\text{network}}(\Phi, \lambda) - F_\times^{\text{network}}(\Phi, \lambda)|^2 d\Omega \right)^{-\frac{1}{2}} \quad (5.1)$$

where  $F^{\text{network}} = \sqrt{\frac{F_1^2 + F_2^2 + \dots + F_{3N}^2}{3N}}$  is the antenna factor, either for the plus polarisation or for the cross polarisation. Since each single detector can be decomposed into 3 interferometers for the future generation GWD (see for example figure 1 of [136]), a  $N$ -detector-network consist  $3N$  interferometers. A larger value of  $I$  represents a network that has a response being more equal in both polarisations.

- $D$ , which measures the accuracy with which the network can localise the sky position of the source.

$$D = \frac{1}{4\pi} \int H(S - A_{90}(\Phi, \lambda)) d\Omega, \quad (5.2)$$

where  $H(x)$  is the Heaviside function,  $S$  is the preset threshold that we will discuss later, and  $A_{90}(\Phi, \lambda)$  is the 90% confidence localisation region for a source located at

sky position  $(\Phi, \lambda)$ . A large value of  $D$  means it's possible to perform good quality sky localisation for larger fraction of the sources.

- $R$ , which measures the accuracy with which the network can reconstruct the parameters of a standard compact binary source.

$$R = \left( \frac{1}{4\pi} \int \sigma_{\mathcal{M}}(\Phi, \lambda)^2 d\Omega \right)^{-\frac{1}{2}}, \quad (5.3)$$

we use the parameter of chirp mass  $\mathcal{M}$  as a standard (*c.f.* chapter 1.2), since it is expected to be the best determined parameter.  $\sigma_{\mathcal{M}}$  is the uncertainty derived from Fisher Information Matrix [17]. A larger  $R$  value represents a network with the ability to perform more accurate parameter estimation.

In the case of the  $D$  metric introduced in [136], we have used a slightly different expression for characterising the accuracy of a GW detector network in source localisation. Our new  $D$  metric is based on the method introduced in Fairhurst 2010 [231], since it gives the localisation error, which is more interesting and has a more direct physical meaning than the previous definition of  $D$ . Here the localisation accuracy for sources at various sky positions is expressed as the angular areas of the 90% confidence localisation regions (ellipses) obtained by triangulation of the source after successful individual detections with  $N$  GW detectors ( $N > 2$ ). For a given  $N$ -detector configuration, we first calculate the localisation ellipses' angular areas, for every given direction, then the  $D$  metric value is computed as the percentage of the sky for which the area of these ellipses falls below a specified threshold,  $S$ . Using the simplification that all detectors in the network register the incoming GW signal with the same timing accuracy, we can directly express  $S$  in  $\text{deg}^2$ . The actual choice of thresholds for  $S$  that we adopted will be explained in detail in 5.2.5. Notice that the calculation of  $D$  should include the timing uncertainty and, according to [231], that value is related to SNR. Here we assume all signals have a timing accuracy that corresponds to an SNR of 8. The influence of this value on  $D$  is degenerate with  $S$  and the actual choice of value could be somewhat arbitrary.

In practice we find out that the new definition of  $D$  is more realistic. However, it is worth mentioning that although the definition of  $D$  has been changed, when considering the optimal configurations, both new and old definitions of  $D$  give very similar results, suggesting an intrinsic consistency between both definitions. So in particular for configurations with high  $D$ , the new definition makes negligible change to the conclusion.

Note also that in what follows we apply equal weights to each of the three individual FoMs. This approach is a gross simplification, and ignores the possibility that the different FoMs may have different scientific purposes and their actual relative importance would depend on the context in which the detector network was operating. For example, in many proposed

applications of GW astronomy a key consideration is to identify an EM counterpart of the GW source in order, e.g., to determine its redshift or some other astrophysical characteristic of the source that is crucial for its exploitation. In this scenario it might be that optimising the sky localisation is the most important of the three FoMs, while in other circumstances it might be that optimising the measurement of the source polarisation (which is important, for example, to break degeneracies in the sky position, or to help detecting un-modelled Burst waveforms) would be preferred. One could straightforwardly adapt our method to such a case simply by adjusting the relative weighting of the FoMs. Furthermore, there may be some intrinsic correlation between the FoMs, which would again call for a more generalised combined statistic that directly takes this into account. We defer such extensions until future work, however, and address here only the case of equally weighted, fully independent, FoMs. However, we recognise that an inappropriate choice of weights could lead to a selection criterion that was not physically motivated.

For each FoM  $I$ ,  $D$  and  $R$ , we find the maximum values  $I_{\max}$ ,  $D_{\max}$  and  $R_{\max}$  from the ideal Earth(see chapter 5.2.5), and normalise the FoM to these maxima. The squared sum of these normalised values will then give the total FoM,  $C$ , as defined in [136]

$$C = \sqrt{\left(\frac{I}{I_{\max}}\right)^2 + \left(\frac{D}{D_{\max}}\right)^2 + \left(\frac{R}{R_{\max}}\right)^2}. \quad (5.4)$$

With such a definition, the total FoM is not biased towards any metric as different scientific goal motivates different figures of merit. For example, the search for burst signal would place more emphasis on good performance of  $I$ , and  $R$  is more important for the search of CBC signals, while  $D$  is equally important for both burst and acCBC.

## 5.2 Description of Method

As briefly discussed in the previous sections, in *Raffai et al.* [136] the network optimisation is achieved by first fixing the other detectors' locations and then finding (using the FoM) the optimal site for an additional detector. Moreover the number of detectors considered in [136] was limited to 3 in a few example cases of a global optimisation. Readers are reminded that the primary motivation for the current work was to extend the earlier results to the case where all detectors could be optimised simultaneously. This would, of course, as a consistency check allow us to compare our results with those of [136], as well as possibly to identify new, optimal sites.

One cannot reliably predict the actual number of GW detectors that will operate in the future since there are so many uncertainties. Once again, we note that we restrict our considerations to only those of future generation detectors. Optimistically, if an unexpected and exciting

new discovery were to occur with the nascent advanced detector network, then – combined with a healthy global economic environment – it seems reasonable that this would boost the case for building several more detectors and pushing their design envelope to the future generation. On the other hand if the actual rate of astronomical events observed by the advanced network is at the pessimistic end of current predictions, then – if it were combined with difficult economic circumstances – this might strongly limit the number of future generation GW detectors that would actually be built. Ideally, therefore, we would want the optimally-chosen sites for proposed future detectors to be as flexible as possible, so that in particular their scientific performance could be high in both optimistic and pessimistic scenarios. Consequently, in this work we also extend our analysis to consider a 5-detector-network, in order to answer the specific question of whether a “good” site for a 3-detector-network is still attractive when a 5-detector-network is considered. Thus we want to determine the optimal location for the first site so that it leaves the future the most flexibility. Of course it is still natural to expect that the optimal sites for a 3-detector-network, supplemented by two additional, optimally-chosen sites, would be at least slightly different from the optimal sites determined simultaneously for a 5-detector-network. However, in the event that not all new detectors would be funded and built at the same time, it may be very likely that a 3- or 4-detector network would be built first and then extended by one or more additional detectors (exactly analogous to the current proposal for LIGO India). Our question about assessing the performance of different detector sites as the size of the network changes would, therefore, seem to be both timely and appropriate.

### 5.2.1 Methodology

It is clear that, in order to optimise all detectors simultaneously while exploring the situation for a network of up to 5 detectors, the method adopted by [136] would not be appropriate. In [136], the geographical regions that are suitable as detector sites were reduced to  $\sim 1,500$  discrete candidate locations and the optimisation was achieved by an exhaustive search over all of these candidates. The computing overhead for this approach is tolerable if one optimises for only one site at a time. However, if we allow even 3 detectors’ locations all to be free parameters over which to be searched, then we have  $\sim 1,500^3 \approx 3.4 \times 10^9$  different combinations to explore. Extending to a 5-detector-network would increase this to  $\sim 7.5 \times 10^{15}$  combinations – which is far beyond what is currently realistic. However, equally clearly, there will be a significant fraction of these combinations that correspond to networks with a relatively low FoM in which we are not really interested. The question then becomes: how can we efficiently explore only those regions with high FoM, even for networks with a large number of detectors? Bayesian inference methods like Markov Chain Monte Carlo (MCMC) [183, 184, 156] or Nested Sampling [166, 167, 157] are designed to deal with such

problems (*c.f.* chapter 2), and they perform especially well when the parameter space has a high dimensionality. We therefore adopt a Bayesian inference approach here, and assign as the posterior some monotonic function of the total FoM.

For an equilateral triangular shaped interferometer, as is being proposed for the future generation GW detector [58], the local antenna response for a given signal is nearly independent of azimuth. Consequently we do not consider the orientation of the detectors in our network, only their geographical location, i.e. longitude and latitude. Thus, if the detector network consists of  $N$  detectors, then the dimensionality of the problem will be  $2N$ . For even a 3-detector network, then, which corresponds to a 6-dimensional parameter space, we can expect that our Bayesian method will significantly outperform the brute force grid-based search.

However, one should realise that the problem we are studying is strictly *not* a Bayesian parameter estimation problem. We are merely taking advantage of some of the technology that has been developed to carry out efficient sampling of high-dimensional Bayesian posterior distributions, and adapting our problem so that this technology may be directly applied to it. Essentially we are only interested in efficiently identifying and sampling from regions with high FoMs. Consequently constraints like *detailed balance*, which in general are required in MCMC applications in order to ensure that the samples are indeed drawn from the appropriate posterior distribution [156], will be of less concern to us here.

There is a vast literature on MCMC methods, including a growing list of example applications in the field of GW astronomy [232, 233, 111]. A simple collection of MCMC methods is illustrated in chapter 2, and the interested reader is referred to the references in that chapter.

As noted earlier, one interesting feature of our approach is that we can shift the entire detector network while keeping fixed the angles between the individual detectors without altering the FoMs that are computed, provided all sites remain within allowable regions. As also noted earlier, including extra information about the actual anisotropic distribution of cosmological sources would break this degeneracy, but such an extension is beyond the scope of the current study.

However, another feature of our optimisation problem is that it is naturally multi-modal, i.e., it contains multiple local maxima. The territory of the Earth is divided into isolated continents and this naturally leads to significant discontinuities in the computed FoMs, thus making the distribution of optimal networks intrinsically multi-modal. The simplest MCMC methods generally become clumsy and less reliable when dealing with posterior distributions that have multiple modes. Hence we have developed an MCMC-based approach that still performs well for multi-modal distributions. More specifically we have taken advantage of MCMC methods' ability to concentrate sampling in regions of high FoM while simultaneously being able to 'swap' between multiple, distinct modes. To meet these require-



ments we have developed a new variant of MCMC, known as *mixed MCMC* (c.f. chapter 4.2), that can sample independently from different regions simultaneously. Other Bayesian sampling methods like parallel tempering MCMC [192, 194, 193], affine invariant MCMC [234, 202, 111] and MultiNest [168, 181] are able to sample from multiple regions, but mixed MCMC is among the most efficient of such methods, as it is designed to be sampling efficiently through multi-modal posterior. We describe fully our mixed MCMC method in chapter 4.2 [235].

### 5.2.2 The actual realisation

The definition of the FoM for a network of detectors given in the previous section is not sufficiently discriminatory from the point of view of implementing MCMC. In particular, from the way in which the equations are set up the FoM for the least optimal network differs by only a factor of two from the most optimal network. If we define the effective ‘posterior’ to be simply proportional to the FoM, therefore, the MCMC sampler will waste a great deal of time in uninteresting regions of the parameter space. Consequently, we manually set the effective ‘posterior’ to be the exponential of the FoM, so that there is much greater differentiation between the least and most optimal networks. This in turn ensures that the sampler will spend more time exploring regions with high FoM.

In order to begin sampling, we manually partition the parameter space in a conservative manner so that one region is allowed to host at most one major mode. The details of how this is done are discussed in 5.2.3. However, we note that the parameter space is not continuous within each region even after partition. In order to avoid any adverse impact of discontinuity, we enable the network to contain sites that are located in unfavored regions like oceans so that the sampler can traverse between discontinuous regions. The exclusion of disfavoured regions is discussed in more detail in chapter 5.2.4. For every such ‘bad location’ site, the posterior will be divided by the base of natural logarithms,  $e$ , so that the sampler does not waste too much time exploring undesired regions. Under this formulation disfavoured regions can therefore be sampled, but are not favoured, and the discontinuity problem disappears. Notice, moreover, that when we present our conclusions about optimal sites, those configurations that contain bad locations will in any case be automatically discarded.

Multiple CPUs were used to sample several network configurations simultaneously, in order to further boost the efficiency. All of the samples generated were combined into 4 groups, and the sample results from these 4 groups were constantly monitored. Once the convergence criteria were met, i.e. the properties of the 4 groups were sufficiently similar, the sampling process was stopped.

We apply an automatic stopping criterion for the convergence of mixed MCMC chain. In

each sub-chain, this convergence criterion is checked using the well-established Gelman-Rubin criterion [185] on every parameter. Interested readers are referred to 5.2.5 for details about the Gelman-Rubin criterion.

### 5.2.3 Constructing the partition

In our original application of mixed MCMC in *Hu et al.* [235], the multiple modes were assumed to have been identified using methods such as parallel tempering, so that the identification is achieved objectively. However, in the application considered here we can further simplify this process by manually partitioning the allowable regions into 6 patches corresponding to the Earth's continents – i.e. North America, South America, Europe, Africa, Asia and Australia, as shown in figure 5.1.

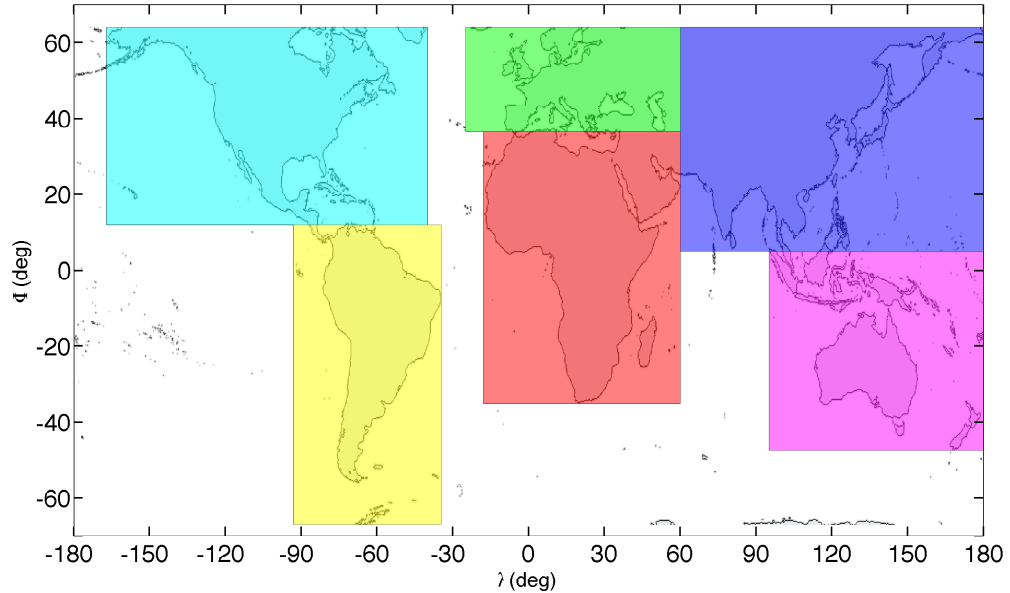


Figure 5.1: Division of the world map into 6 distinct sections, roughly speaking coincident with the 6 populated continents: North America, South America, Europe, Africa, Asia and Australia. Notice that, although there are some areas that are not covered by any section, these areas are mostly islands and consequently all will be rejected either by the criterion of lying in a polar region or the criterion of being less than 100 km inland.

We assume that for  $n$  detectors, each of them can be located in one of these 6 continents. If one continent can only host up to one detector, then the number of different possible combinations is just simply  $C_6^n$ . In the case of a 3-detector network, this number is  $C_6^3 = 20$ . However, we should also consider that some continents like Asia are very large in area, so that locations in one continent can be a considerable distance apart. In order to be conservative, therefore, we enable each continent to host more than one detector. Hence we must

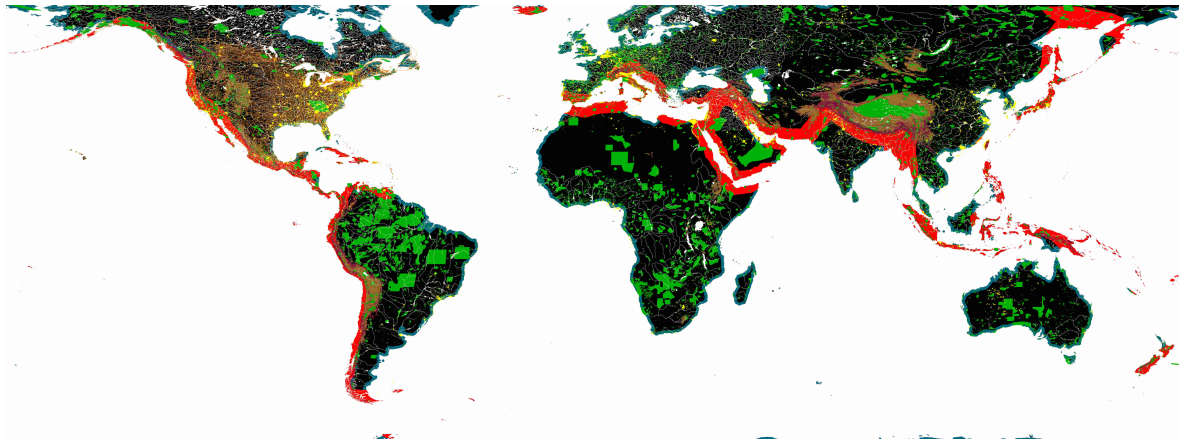
include among the possible configurations those networks for which one continent hosts two detectors; in total this makes  $C_6^1 C_5^1 = 30$  cases, together with the  $C_6^1 = 6$  cases in which all detectors are located in one continent. So for the 3-detector network, there are in total 56 different possible combinations of host continents; in each such combination, one mode of mixed MCMC is assigned to sample and we identify combinations with modes. Analysis of the 5-detector network situation is identical in principle, although somewhat more complicated: it is straightforward to show that there are in total 252 different combinations in this case.

Notice the assumption that is implied here, that there will be generally no more than 1 local posterior peak in one mode. We believe that this assumption is sound and justified since we expect the figures of merit to be slowly changing, and it can help us vastly to simplify the process of initialising the chains.

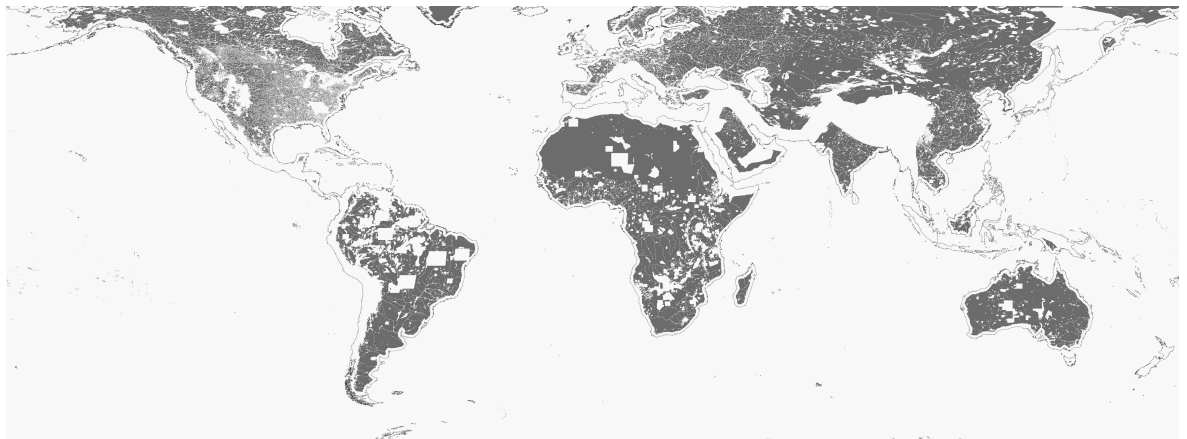
#### 5.2.4 Exclusion of unsuitable regions

In order to filter out regions where it would be unsuitable to build a detector, criteria similar to those in [136] were used to exclude such sites. First, to build a detector underwater would be far from realistic due to the huge expense to build and maintain it, so we exclude all oceans, seas and continental lakes. Furthermore, we exclude all coastlines that are within  $\sim 100$  km distance from the ocean, so that the micro-seismic noise due to oceanic waves is mitigated. Similar considerations of transport and the convenience of maintenance lead to the exclusion of polar regions, regions with slope steeper than  $5^\circ$ , as well as regions with an elevation higher than 2000 meters above sea level [236]. Also, routine human activity in centres of population would induce a gravity gradient noise, so the detectors should be built far away from densely populated regions, including major roads in North America. This exclusion is achieved by excluding regions where there is significant artificial illumination during the night, in addition to populated areas as defined according to the Natural Earth database [237, 238]. In addition, and differently from [136], we further exclude protected areas, since undertaking construction in such regions would be generally illegal [239].

This work used the exclusions as mentioned above. For future work, more constraints have been proposed, such as seismically unstable regions which are not suitable because of their high level of environmental noise [240]. We show our most up-to-date exclusion figure in 5.2. Beyond that, more factors can be considered, like military regions, which might not be feasible to access; regions with active mining activity, or which contain rich mines, that could induce gravity gradient noise to the detector data. Contaminated areas and nuclear test sites are also not ideal places.



(a)



(b)

Figure 5.2: The figures show the exclusion of unsuitable regions according to various exclusion criteria, including a coloured version and a monochromatic version. Interested readers are encouraged to check out the full resolution files at: <http://elysium.elte.hu/~praffai/geomap.html>. Detailed conditions and corresponding colors (for the coloured version) are listed below.

5.2a: Dark green: coastlines with  $\sim 100$  km width. Red: seismically active areas with a 200 km width [240]. White: oceans, seas and fresh waters, according to the Natural Earth database [237]. Dark brown: roads of North America [237]. Brown: elevated areas, *i.e.* areas above 2000 m, according to ASTER GDEM Worldwide Elevation Data map [236]. Claret: high gradient areas, *i.e.* areas with higher than  $5^\circ$  slope [236]. Green: protected areas like national parks, according to the World Database on Protected Areas [239]. Yellow: populated areas, including bright areas on the NASA night lights map [238] and other populated areas like those in India and China [237]. Black: potentially suitable regions.

5.2b: White: outfiltered regions. Grey: potentially suitable regions. The coastlines are marked with a solid black contour.

### 5.2.5 Optimising parameters

Originally, the calculation of FoM for one single network configuration would take  $\sim 100$ s – clearly emphasising the need for optimisation of a conventional MCMC requiring of order  $10^5$  samples.

First, the original calculation of the  $I$  and  $R$  FoM was achieved through numerically integrating over the whole sky with a very high angular resolution. We then tested empirically the relation between sky resolution and calculation accuracy using a detector network on an hypothetical idealised Earth with the first detector site fixed to be at the North Pole, the second set fixed along the Prime Meridian of longitude and the future site placed uniformly over the surface of the sphere. Because of the symmetry inherent in our FoM, any actual multi-detector network can be rendered equivalent to this idealised network through appropriate choice of coordinate system.

To calculate  $I$  or  $R$ , we first discretise the whole sky into uniformly distributed representative points using the *healpix* algorithm [241], and calculate corresponding individual FoMs for a source located at each of these points. The input to *healpix* is a positive integer that determines the resolution: the higher this input parameter, the more points we discretise and the more accurate  $I$  and  $R$  will be – although the calculations will also be more time-consuming. Since the actual number of calculations is proportional to the square of the resolution, the gain in computational efficiency from reducing this *healpix* parameter is huge. We calculated the individual  $I$  and  $R$  FoMs for a range of different resolutions, and compared them with the values obtained using the highest resolution that was feasible – which we defined as that obtained for a *healpix* parameter of 8. The relative difference in the FoMs, averaged over all combinations of networks from the aforementioned hypothetical Earth, was then calculated for smaller values of the *healpix* input parameter. As shown in figure 5.3, a much lower resolution (with a *healpix* parameter of 3) will only result in a negligible loss of accuracy in the average FoMs. When the resolution is increased, variance is expected to increase, causing the non-monotonic behaviour seen in figure 5.3. We therefore adopted a *healpix* parameter of 3 in our subsequent analysis, corresponding to an angular resolution of 0.13 radian, or 7 degrees. We understand this choice as a natural outcome of the fact that 7 degrees is much smaller than the characteristic angular length for the antenna pattern to vary.

Secondly we explored the optimal number of CPUs to use. It is natural to expect that using more CPUs should lead to greater computational efficiency, and this was one of the main motivations for the development of mixed MCMC. However, we can expect that there will be a limiting behaviour such that, for a sufficiently large number of CPUs, further increasing this number will not result in any significant further improvement, since every sampler needs some certain sample time to get some reliable estimation of the total parameter space. We illustrate this trend in figure 5.4, where we see that the improvement in efficiency reduces

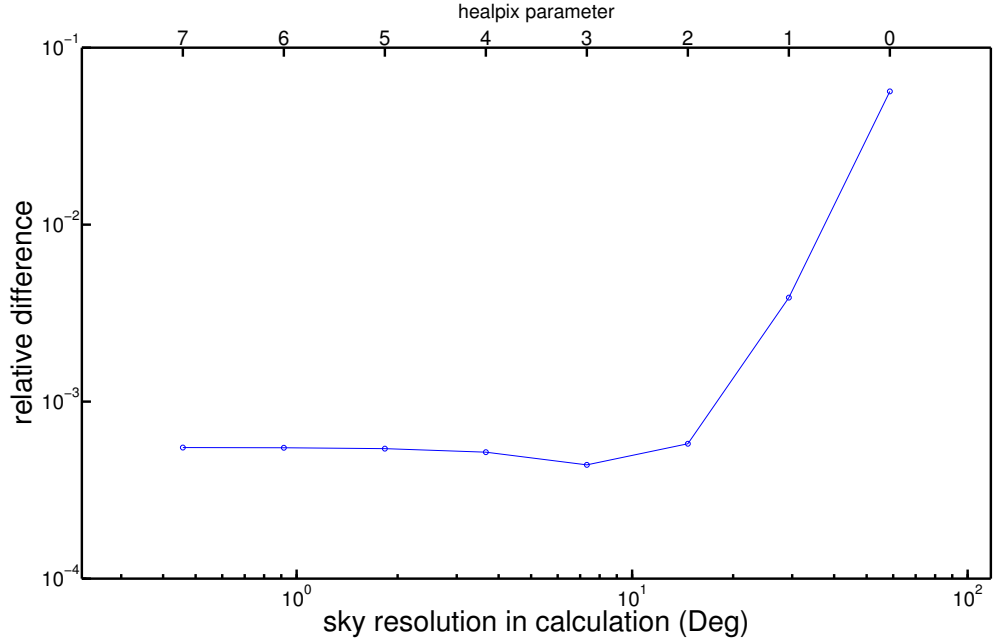


Figure 5.3: Relative change of the  $I$  FoM, averaged over all possible combinations from the ‘hypothetical Earth’ as described in the text, as a function of *healpix* input parameter. As expected, when the resolution is made smaller, the relative difference becomes smaller, while the uncertainty also becomes smaller, which explains the minimum value in the *healpix* parameter 3.

substantially after the number of CPUs exceeds  $\sim 40$ . Thus we determined the optimal number of CPUs to be 40 when sampling for a 3-detector network. We assumed further that this optimal number should be proportional to the total number of distinct network configurations, as described in 5.2.3. Hence for a 5-detector network we adopted as the optimal choice of CPU number  $40 \times 252/56 \sim 200$ . Our numerical test runs showed that for a 5-detector network, using 200 CPUs did indeed sample much faster than using 40 CPUs, while still yielding satisfactory results.

Also, we investigated the usage of Gelman-Rubin criterion [185]. When multiple realisations of the same model are running simultaneously, the properties of MCMC guarantee that different realisations should give a similar distribution after a sufficiently long time. The variance of each parameter within each MCMC realisation was calculated, and the estimation of the intrinsic variance was constructed with the information of multiple realisations. The ratio between this estimate of the intrinsic variance and the variance within each realisation is thus calculated; this is defined as the Gelman-Rubin  $R$  value. Due to the way it is constructed, the  $R$  value on every parameter always tends to be larger than unity, but when the MCMC chain has converged it will asymptotically tend towards  $R = 1$ .

In this problem, we set a Gelman-Rubin criterion of  $R = 1.1$  – that is, if no parameter in a given subchain yields a Gelman-Rubin  $R$  value larger than 1.1, we will label this subchain as

“converged”. However, the sampling in this sub-chain is continued, otherwise the property of detailed balance will be totally destroyed and we can’t predict its impact on our sampling result. In order to compute the Gelman-Rubin  $R$  for  $m$  different MCMC realisations of  $n$  points  $x_i$ , one needs to construct

$$B = n \sum_{i=1}^m (\bar{x}_i - \bar{x})^2 / (m - 1) \quad (5.5)$$

where  $\bar{x}_i$  and  $\bar{x}$  are the mean of each MCMC and total sample, separately.

$$W = \sum_{i=1}^m s_i^2 / m \quad (5.6)$$

is the average of the variances defined as  $s_i^2$ . Based on these values, one can compute

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{B}{n} \quad (5.7)$$

$$\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^2 + B/mn} \quad (5.8)$$

$$dof = 2\hat{V}^2 / \hat{v}ar(\hat{V}) \quad (5.9)$$

for the estimation of the target variance, the estimation of the sampling variance and the number of degrees of freedom respectively, with

$$\hat{v}ar(\hat{V}) = \left(\frac{n-1}{n}\right)^2 \frac{1}{m} \hat{v}ar(s_i^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1} B^2 \quad (5.10)$$

$$+ 2 \frac{(m+1)(n-1)}{mn^2} \frac{n}{m} [\hat{c}ov(s_i^2, \bar{x}_i^2) - 2\bar{x} \hat{c}ov(s_i^2, \bar{x}_i)] \quad (5.11)$$

and the  $R$  value is defined as  $\sqrt{R} = \sqrt{(\bar{V}/W)dof/(dof-2)}$ . As  $dof$  tends to infinity, the term  $dof/(dof-2)$  will cancel out for large  $n$ . So  $R-1 \sim \frac{m+1}{m-1} \frac{\sum_{i=1}^m (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^m s_i^2}$  [185].

Note that in Bayesian parameter estimation, a conventional choice for the threshold to be converged is  $R-1 \leq 0.01$ . However, in our case, in order to accelerate the sampling process we run in parallel  $4 \times N$  chains and then sum the  $N$  chains up to make up 4 “major chains”. The Gelman-Rubin  $R$  criterion is then applied to these 4 major chains and a relatively large criterion  $R-1 \leq 0.1$  is adopted. So for 3-detector-network, the stop criterion translates as meaning that the variance of average should not be larger than 0.08 times of the average variance within groups, and for 5-detector-networks 0.09 times.

The sampling process is then continued until all subchains have converged. We expect that the results would not be changed significantly if we were to adopt a stronger constraint on the  $R$  criterion. However, smaller  $R$  criterion value would largely increase the computational cost. The choice of  $R = 1.1$  is an sufficiently tight constraint while resulting in an acceptable

computational burden.

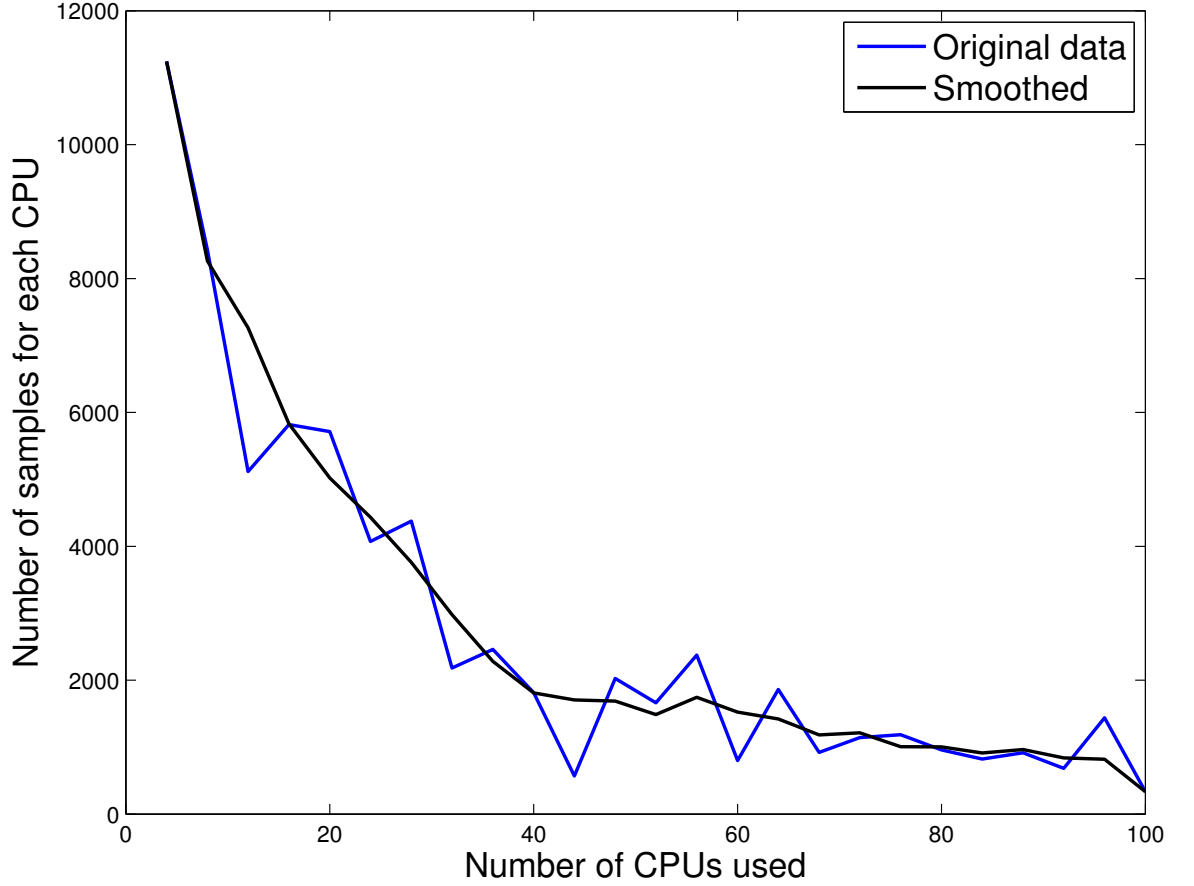


Figure 5.4: Average number of samples for a 3-detector network using different numbers of CPUs. Since the Gelman-Rubin criterion was applied on 4 major chains, the tested numbers are multiples of 4 (for details see section 5.2.2). We can see that there is a uniform decrease in the number of samples as more CPUs are used. We apply a moving average filter with 5 points to obtain the smoothed data. The vertical axis is just for illustration, and does not correspond to the actual number of samples obtained, since we apply different stopping criteria in this test compared with that used in our actual sampling. However our conclusion about the optimal number of CPUs suitable for our analysis should remain valid.

Finally consider the choice of value of  $S$ , which is required to calculate  $D$ . Notice that according to [231], sky localisation error is strongly related to SNR, which for a detection criterion is chosen to be 8 in general cases. In previous studies, a typical event’s localisation error will span an area of  $10 - 100\text{deg}^2$  with 2-3 GW detectors, and  $\sim 5\text{deg}^2$  for a 5 detector network (e.g [231]). Besides, the rule of thumb for choosing proper  $S$  is to make the FoM is as discriminatory as possible. If we set  $S$  too small, then most networks will give a value barely larger than zero, and the influence of calculation uncertainty will be severe. If, however,  $S$  is too large then in almost all networks the fraction of the sky for which a source can be localised to a region smaller than  $S$  also becomes very large, and once



again we will lose the ability of the FoM to distinguish effectively between different network configurations.

For the 3-detector network, we find that  $S$  is best set to  $\approx 59.5 \text{ deg}^2$ , in this case, the best configuration gives a  $D$  close to but not equal to 1, thus avoiding the aforementioned degeneracy. For the 5-detector network, however,  $S$  is set to the much smaller value of  $S = 2.5$  area units, translating to  $4.5 \text{ deg}^2$ . This is not surprising since it is generally expected that networks with more detectors will perform better with regard to this FoM. Tests on an ideal Earth, as described in the previous sub-section, show that adopting the value of  $S = 2$  will lead to a significant degeneracy at  $D = 0$ , while a value of  $S = 3$  leads to severe degeneracy at  $D = 1$ . Hence the choice of  $S = 2.5$  area units represents an appropriate trade-off between these two cases, which is equivalent to  $4.5 \text{ deg}^2$ . These values shows good consistency with [231].

Notice that the ability of sky localisation is closely related to the EM follow-up. Since the sky localisation area tends to be much bigger than ordinary telescopes' fields of view, one would be much more interested in the value of sky localisation area, but not so interested in the shape or topology of the localisation. Thus we have good reason to believe that our definition of  $D$  is reasonable and representative.

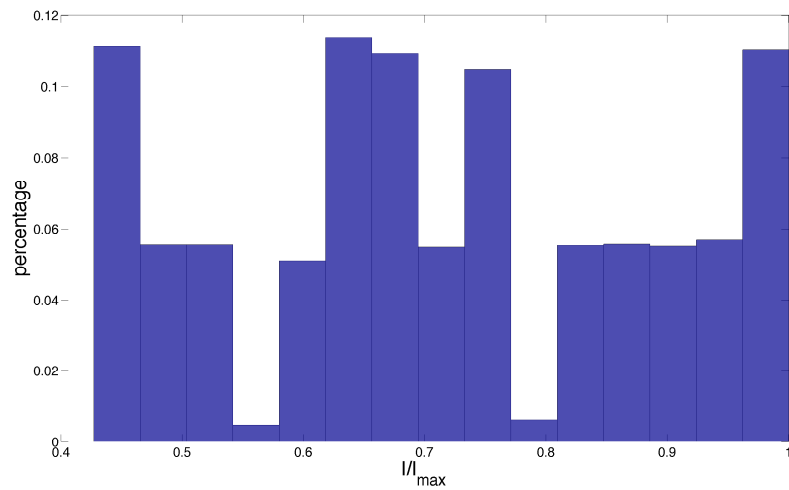
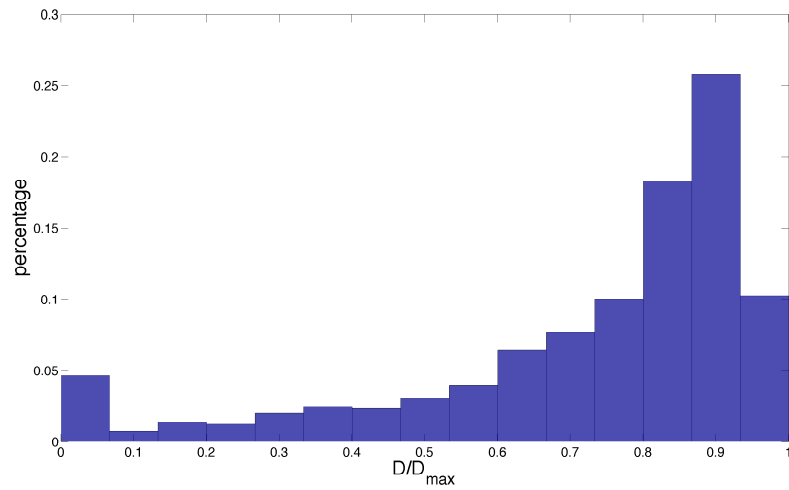
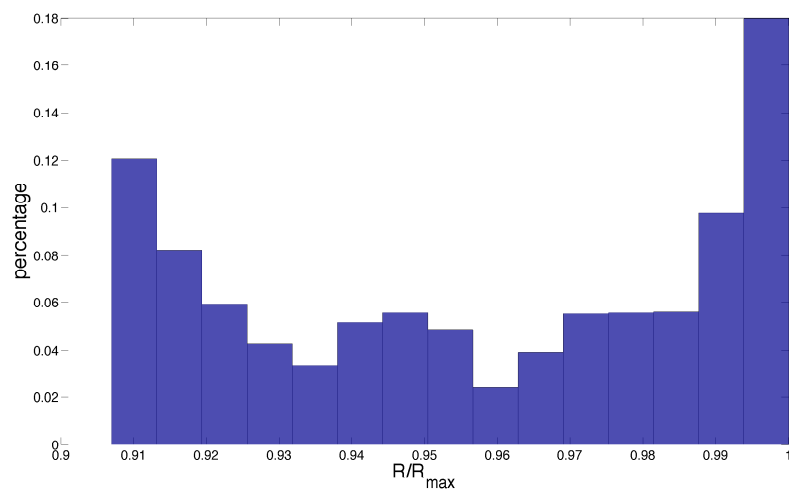
Based on the result from the ideal Earth, we show the normalised histogram of three figures of merits individually in figure 5.5a - ??.

## 5.3 Results

The principal goal of this work is to determine the optimal sites on which to build a network of future generation GW detectors. However, as noted earlier, the formulation of our FoM is invariant under translations of the entire network across the Earth's surface, provided that the network shape remains unchanged and all sites remain located in allowable regions.

In order to better distinguish good sites and good network configurations from others, we can define the “*flexibility index*” of a site – *i.e.*, if that site is included in our network how many possible distinct network configurations could there be that would each give a high FoM? In this sense, a “better” site means one that would give more freedom, or flexibility, in choosing the locations of other detector(s). Our results below, therefore, illustrate on a world map the flexibility index of different sites for networks containing different numbers of detectors.

We run our optimisation code separately for 3-detector and 5-detector networks. The mixed MCMC method is used so that the sampler can be concentrated in separate regions of high FoM. Once the sampling is terminated, when the convergence criteria are met, we assume that all regions of interest should by then have accumulated sufficient samples to represent

(a) Normalised histogram for the  $I$  on the ideal Earth.(b) Normalised histogram for the  $D$  on the ideal Earth.(c) Normalised histogram for the  $R$  on the ideal Earth.Figure 5.5: Individual normalised histogram of  $I$ ,  $D$  and  $R$ , conclusion drawn from the ideal Earth simulation.

adequately the underlying distribution. Although we want the maps to reflect the globally optimal network configurations, for networks containing different numbers of detectors, we should bear in mind the other external factors that will determine in reality where future generations of detectors are constructed. Thus we adopt a threshold on the FoM, 90% of the highest FoM, and consider *all* networks that exceed this threshold. We then determine our flexibility index for each site by counting the number of different networks including that site which exceed our chosen FoM threshold.

For displaying our results we adopt a standard world map with  $1520 \times 759$  pixels. For each pixel, we convert the pixel location into geographical coordinates, and identify all the networks that contain one site within a certain distance ( $\sim 200$  km for the zoom-out figures like Figure 5.6 and 5.9, and  $\sim 64$  km for the zoom-in figures like Figure 5.7, 5.8 and 5.10) of that pixel. This choice leads to a larger fluctuation for the zoom-in maps.

To avoid multiple counting of ‘similar’ configurations, we determine the inner product of the unit vectors constructed from the coordinates of the sites that comprise two networks that we wish to compare. For a network that consists of  $t$  detectors, we set  $\mathbf{N} = (\Phi_k, \lambda_k, \dots, \Phi_l, \lambda_l)$ , while  $k, \dots, l$  is permuted over  $1, \dots, t$ . The normalised vector  $\mathbf{n} = \frac{\mathbf{N}}{|\mathbf{N}|}$ , and we define the inner product of two networks  $\mathbf{n}$  and  $\mathbf{m}$  to be  $s = \max(\mathbf{n} \cdot \mathbf{m})$  that has been maximised over permutation of component detectors for  $\mathbf{n}$  and for  $\mathbf{m}$ . We define the two networks as duplicates if their inner product is larger than 0.95; in this case one of the network configurations is discarded. The adoption of this approach ensures that our results are not adversely affected by the somewhat arbitrary partitioning of the world map, while at the same time allowing our mixed MCMC approach to benefit from the fast identification of regions of interest precisely through use of this partitioning. Since this inner product criterion of 0.95 is applied to unit vectors, it is independent of the number of detectors in our network. This is very useful as it allows us to compare directly and straightforwardly our results for 3-detector and 5-detector-networks.

To sum up, instead of looking at the sample density in a 6 or 10 dimension parameter space, we convert our result back to a two dimensional Earth surface, by introducing the idea of “flexibility index”. This is calculated by find the number of good networks which contains one component around an interesting location. There are redundancy in the samples as there are some networks which are very similar to each other, and we reduce the networks into representatives, so that a fair comparison could be drawn.

### 5.3.1 Network of 3 detectors

As shown in figure 5.6a, after filtering out all configurations that have a FoM smaller than 90% of the highest FoM, the ‘best’ site (as defined by the flexibility index introduced in the

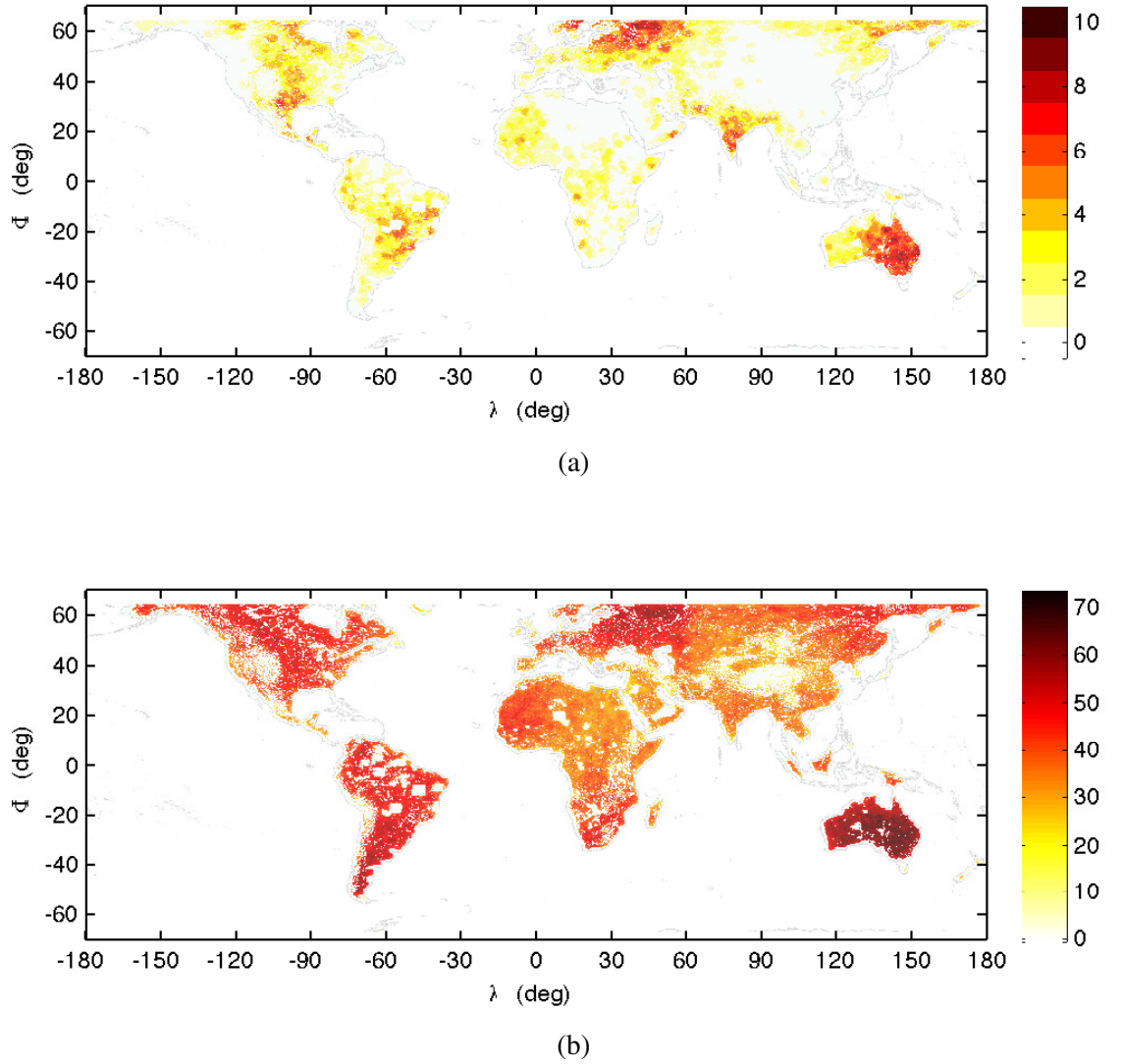


Figure 5.6: World map showing the “flexibility index” – i.e. the number of different, distinct network configurations associated with a particular site, assuming a 3-detector GW detector network. The upper panel shows the result after filtering out all configurations with FoM smaller than 90% of the most optimal configuration, while the lower panel shows the result with the filtering criteria set to be 80%.

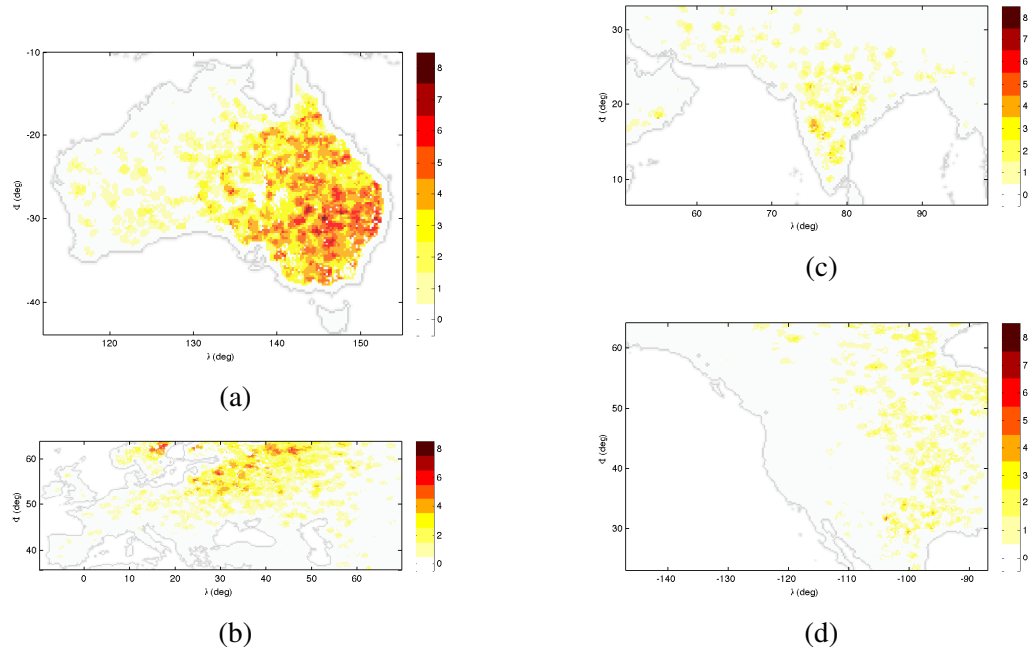


Figure 5.7: Zoom in map for regions with high flexibility indices, from figure 5.6a. Subplot a, b, c and d shows maps of Australia, Europe, India and America respectively. Notice that a shorter smooth length is applied here compared with figure 5.6, causing a larger fluctuation.

previous section) is located in Australia. This result is consistent with previous conclusions from [136]. For the best site identified in this way there are in total 10 different possible network configurations. Notice that in some regions of Europe, North America and India there are also large numbers of alternative network configurations with high FoM, and Australia is only slightly better than these regions.

Figures 5.7a to 5.7d show zoomed-in detail on those regions around the world which have highest flexibility index: Australia, Europe, India and North America.

We also applied a brute force search on the ‘ideal Earth’ (see 5.2.5) in order to check the optimal network configuration. This exhaustive method can search for the optimal configuration of a 3-detector network when not including any terrestrial constraints. We find that such a network would be optimal when the detectors form an isosceles triangle with two sides of length of  $\sim 130^\circ$  on the Earth surface, and the distinct third side of length  $\sim 50^\circ$ . This result further confirms the previous result that when a network consists of only two detectors, the optimal situation is when they are separated by  $\sim 130^\circ$  [136].

We can also notice that Central Africa and East Asia seem not to be ideal sites. The major reason for this is that the  $130^\circ$  circle around these regions mostly falls in the ocean.

In figure 5.6b, the world map is shown for the case of a lower FoM threshold: here we have filtered out network configurations that have FoM less than 80% of the optimal value. Notice that some areas in Africa and in East Asia that were blank in the previous figure, with a

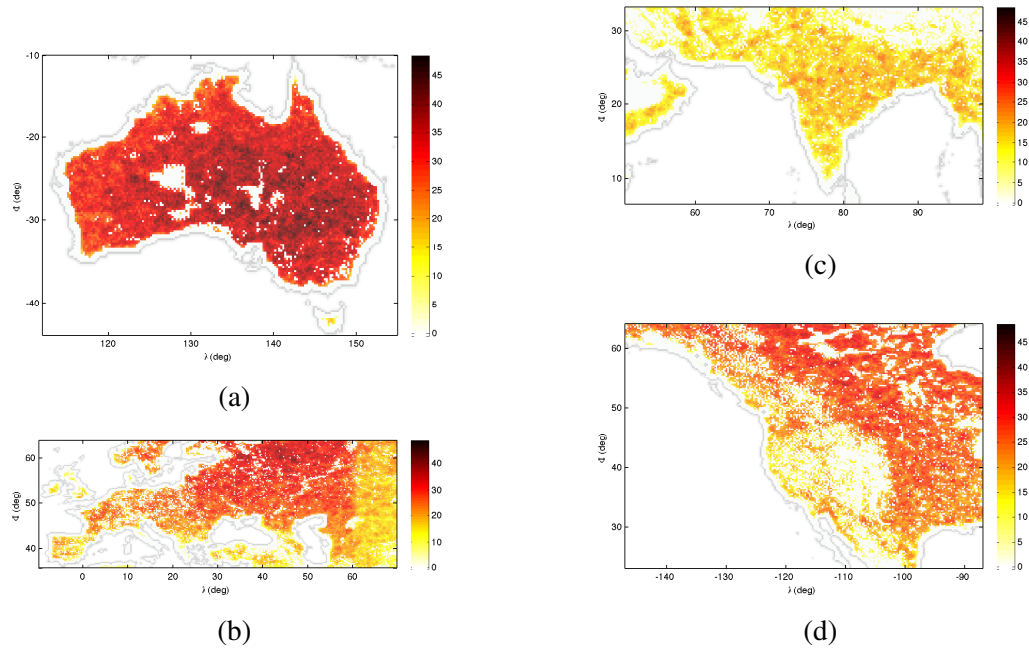


Figure 5.8: Zoom in map for regions in a 3-detector network with high flexibility indices, from figure 5.6b. Subplot a, b, c and d shows maps of Australia, Europe, India and America respectively. Notice that a shorter smoothing length is applied here compared with figure 5.6, causing a larger fluctuation.

90% threshold, are now filled in. This indicates that the blank regions in figure 5.6a were not the result of insufficient sampling but rather were due to the intrinsic lack of high FoM configurations in these regions. It seems that these new potential interferometer sites are either too close to the sea, or are otherwise not ideal for building future generation detectors, however.

In table 5.1 we list the locations of the other two sites for various “good” 3-detector networks, in which one site is located near to the global optimal site.

### 5.3.2 Network of 5 detectors

Figure 5.9 shows a world map of the flexibility index for a 5-detector network. Here we find that Australia is still the best site, and unlike the situation with a 3-detector network, it is significantly better than any other regions. In the best site, there are in total 131 different possible network configurations after filtering out all configurations with FoM smaller than 90% of the optimal FoM. Lowering the threshold to 80% increases this number to 235. Besides Australia, the next best site locations are in North America and South America, although the flexibility index for these locations is more than 50% smaller than for Australia. Another important respect in which our 5-network results differ from those of the 3-detector network is that the first detector can be built almost anywhere, as long as it is not excluded

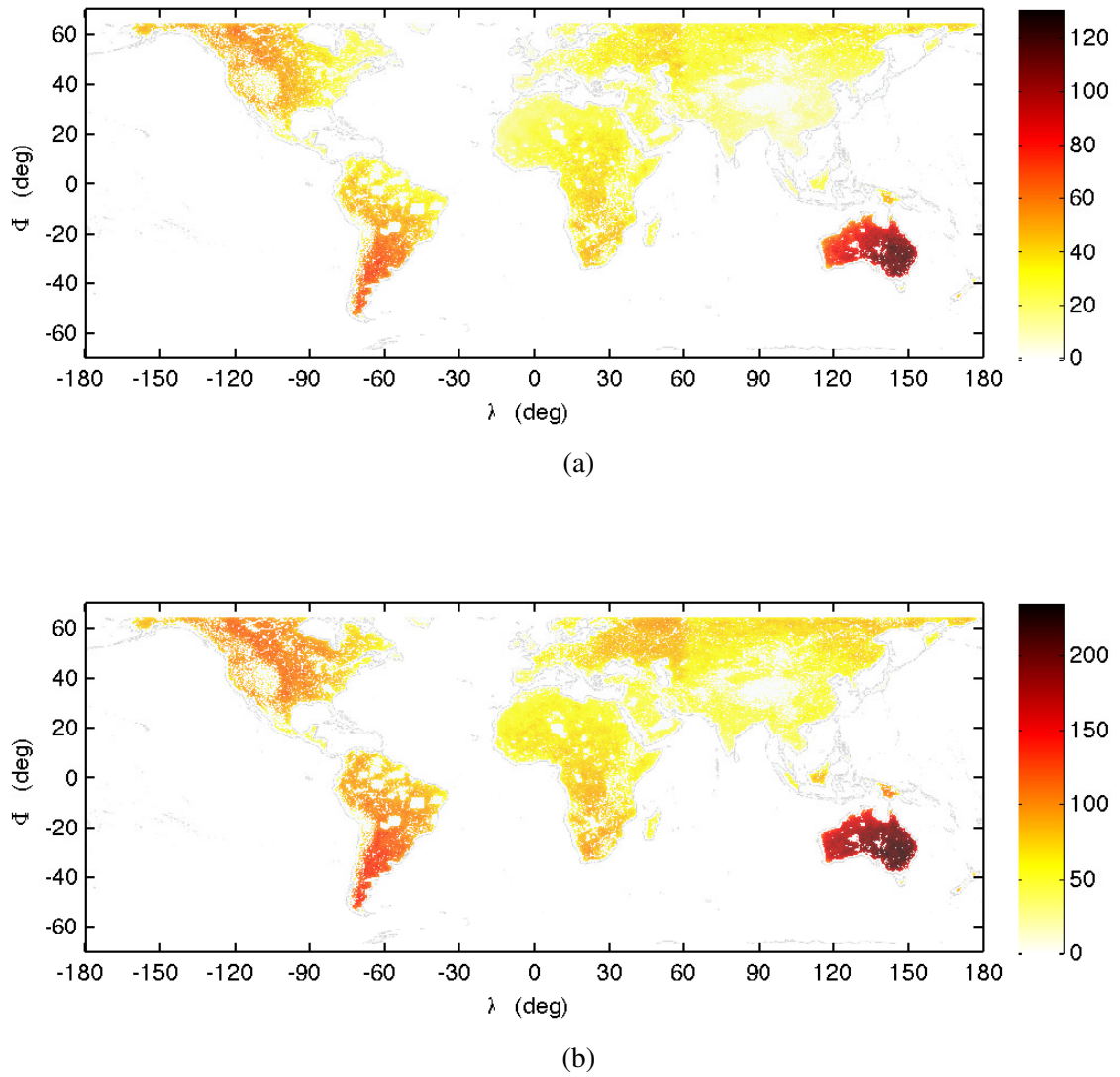


Figure 5.9: World map showing the flexibility index, assuming a 5-detector GW detector network. The upper panel shows the result after filtering out all configurations with FoM smaller than 90% of the optimal configuration, while the lower panel shows the result with the filtering criteria to be 80%.

Site 1		Site 2	
long(°)	lat(°)	long(°)	lat(°)
-84	14	-43	-15
27	51	-99	62
-105	57	-76	1
25	1	-103	37
27	20	17	64
53	62	-69	-16
117	3	-92	47
16	-3	-43	-20
15	-8	21	45

Table 5.1: The locations of the other two detectors, for a series of examples of “good” 3-detector networks, in which the future generation detector is located around the global optimal site, at longitude  $146^\circ$  and latitude  $-30^\circ$ . Each of these networks yields a value of  $C$  that is greater than 90% of  $C_{max}$ . The order does not have any special meaning, so the site 1 and site 2 are interchangeable.

by the conditions described in chapter 5.2.4. One should not be surprised by this outcome since the more detectors that a network includes the more flexible it should become.

Figures 5.10a to 5.10c show zoomed-in detail on those regions around the world which have highest flexibility index for a 5-detector network, and adopting a FoM filtering threshold of 80%.

In table 5.2 we list the locations of the other four sites for various “good” 5-detector networks, in which one site is located near to the global optimal site.

## 5.4 Conclusions and Discussion

In this work we have applied a novel method for identifying optimal sites for future networks of Gravitational Wave detectors. Our method adopted a new sampling approach that is well-suited to dealing with high-dimensional parameter spaces, thus permitting for the first time the simultaneous optimisation of parameters for multi-detector networks – a significant extension of the method previously presented in [136]. We presented results for networks comprising 3 and 5 GW detectors.

We adopted a FoM for each network configuration based on its capability of reconstructing the signal polarisation, accuracy of source localisation and accuracy of source parameters estimation for a standard compact binary source. We followed the definition previously adopted in [136] for the combined FoM,  $C$ ; however, the actual definition of  $D$  in this work is slightly different, as we calculate the fraction of sky for which the source can be localised to better than a specified area,  $S$ .



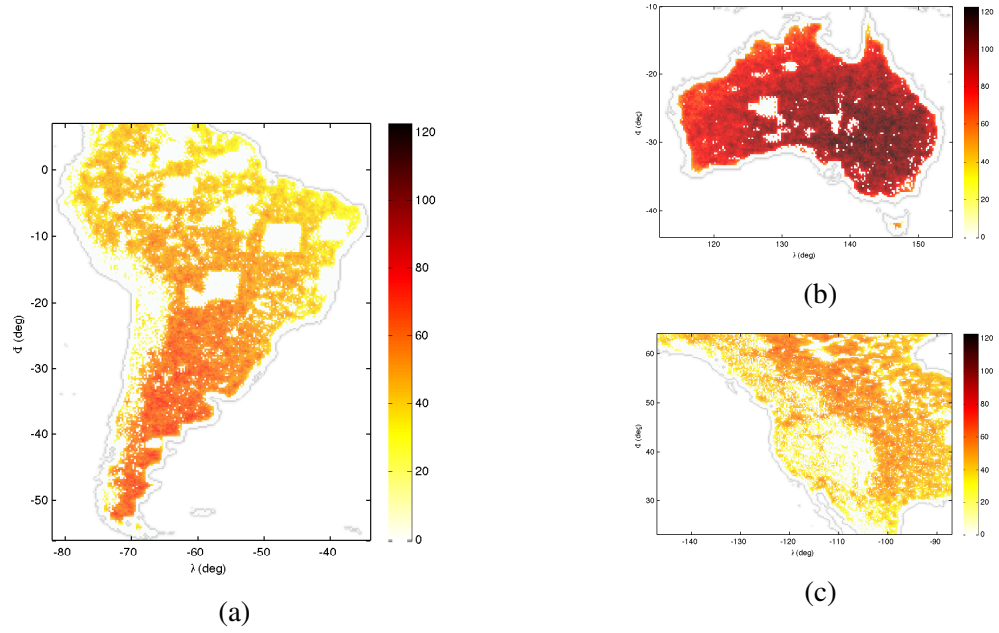


Figure 5.10: Zoom in map for regions in a 5-detector network with high flexibility indices, from figure 5.9b. Subplot a, b, and c shows maps of South America, Australia and America respectively. Notice that a shorter smoothing length is applied here compared with figure 5.9, causing a larger fluctuation.

Site 1		Site 2		Site 3		Site 4	
long(°)	lat(°)	long(°)	lat(°)	long(°)	lat(°)	long(°)	lat(°)
18	-4	32	59	-109	30	-71	-49
23	-29	109	61	74	25	-93	44
19	-2	114	61	-131	58	-64	-5
21	-29	42	62	107	61	-73	5
24	-9	117	41	-118	37	-70	-42
32	-7	74	49	-88	38	-69	-15
35	-11	46	24	-119	41	-63	-36
4	32	19	-23	-114	40	-64	-36
20	-8	-122	44	-71	-45	-57	-13
23	-31	19	51	150	61	-95	30

Table 5.2: The locations of the other four detectors, for a series of examples of “good” 5-detector networks, in which the 5<sup>th</sup> detector is located around the global optimal site, at longitude 146° and latitude −29°. These examples are the networks that give the 10 largest sampled values of  $C$ . The order does not have any special meaning, so the four sites in one line are interchangeable.

We used a MCMC-based sampling method which meets our requirement to sample efficiently in high-dimensional parameter spaces. However, for a multi-modal posterior distribution standard MCMC is much less effective. We have, therefore, developed a variant, known as mixed MCMC, that is ideally suited to handle the multi-modal feature of our problem. We partition the parameter space manually to facilitate fast initial sampling, dividing the world into 6 regions, roughly overlapping with the normal definition of the continents. The multiple modes of our distribution are expected to be located distinctly in combinations of these 6 regions. Such a partition is somewhat arbitrary, but this is intuitive, and the purpose is to distinguish all modes so that no two modes share one piece of the partition. As long as the partition is dense enough, we should obtain conclusions that are robust against changes in the partition.

The sampling results were combined and for each pixel in a  $1520 \times 759$  pixel world map (corresponding to a resolution of approximately 26 km at the equator), we counted the number of distinct network configurations that have a FoM higher than 90% of the best FoM identified. We call this number the “flexibility index” of the network: in this sense, a site with a large flexibility index offers more options for network configurations with a high FoM. In other words, once a detector is built on such a site, one has greater flexibility for locating other detectors. This criterion to identify a “good” detector location is, therefore, well suited to the (likely) case where future detector networks are not built simultaneously but sequentially.

For both 3-detector and 5-detector networks, we consistently found that Australia hosted the best site – further confirming and generalising the conclusions of previous work in [136], where only one detector was optimised. However, for the 3-detector network, the best sites in Australia are only slightly better (in terms of their flexibility index) than optimal sites in Europe, America and India. For a 5-detector network, on the other hand, Australia is a considerably better site than any other region. This would suggest that, if the long-term goal is to create a network of as many as 5 GW detectors, then building one of the first detectors in Australia is a powerful strategy.

We have included two tables (in table 5.1 and in table 5.2) showing the example locations of sites – in networks of 3 GW detectors and 5 GW detectors respectively – which, when combined with a detector located at the global optimal site, yield a combined FoM  $C$  that is more than 90% of  $C_{max}$ .

Our approach is simplified in several important respects – not least our assumption that the spatial distribution of GW sources is isotropic, and the seismic environment is homogeneous within the allowable regions, so that the comparison of different networks is purely determined by their relative geometric shape and (provided that all sites remain within allowed regions on the Earth’s surface) is insensitive to translation of the entire network configuration. The definition of the FoM could be improved to include such factors as economic

stability and scientific policies. Given the difficulty in modelling them accurately, especially over a timescale of decades, we have not considered such factors in this work. The final decision of such a site selection would have to account for them. Nonetheless, we worked out a solid framework that allow future updates and the inclusion of any emerging geopolitical, military, financial, etc. constraints. We are planning to create a crowdsourced project to update and refine the constraint map with all available information, so that we can keep the boundary conditions up-to-date and detailed.

In the future, we can include phase information of GW waveform into consideration; arrange the weights of individual FoMs to be unbiased as some FoM could be more sensitive to configurations than others; correlation between FoMs should be investigated, since for future generation detectors, the ability of distinguish polarisation could be used to help constraining sky localisation [242, 243]. In future work we will also extend our approach to include astronomical information about the actual, anisotropic, distribution of potential GW sources on the sky; this information will break the degeneracy of our FoM to network translation. Meanwhile, seismic stability for a site can also be considered quantitatively in the FoM of the network, further providing more realistic optimisation. <sup>1</sup>

---

<sup>1</sup>This work is accepted by *Classical and Quantum Gravity*, and I am the first author as well as the corresponding author.

## Chapter 6

# The Significance Mock Data Challenge

In this section I will present my work collaborated with some other researchers. The effort is to set up a Mock Data Challenge for the estimation of significance, and let participants analyse and feedback the analysis for the false alarm probabilities, so that we can understand the effect of removal of the zero-lag events in the estimation of false alarm probability. I was responsible for the generating of such Mock Data, and I analysed the feedback to summarize and form the final conclusion.

The ultimate purpose of the work of this thesis, as well as more broadly the work of the entire LIGO Scientific Collaboration (LSC), is to make the direct detection of gravitational waves, only after which the detailed study of gravitational-wave physics and astronomy is possible. In order to claim the first detection confidently, one needs to understand the False Alarm Probability (FAP) of the events well. Such understanding is vital especially for the first detections when they are generally believed to have relatively low False Alarm Probability (FAP), it's harder to make the decision of whether the detection corresponds to a foreground signal or a background noise. However, there exist debates over the proper way of estimating such FAP, which generally diverge from each other in the case of interesting significant events. Thus, the author of this thesis together with Chris Messenger and Martin Hendry set up the significance Mock Data Challenge, also known as the “Hamlet test”, aiming to determine the proper method of estimating FAP. Participants of this MDC include Thomas Dent, Collin Capano, Chad Hanna and Matt West.

The global network of advanced GW detectors is poised to make its first direct detection [56, 52, 244]. The most likely astrophysical sources for detectable transient signals are inspirals and mergers of compact stellar binaries: binary neutron stars, black holes, or mixed binaries [79]. Detection of compact binary coalescence (CBC) would open a new window of gravitational wave astronomy [2]. Upon first, and subsequent detections, the LSC (LIGO

Scientific Collaboration) and Virgo collaborations (LVC) will be tasked with assigning a measure of statistical confidence to the events in question [149]. This FAP estimate will allow GW data analysts to distinguish between sources of astrophysical origin and non-astrophysical detector noise artefacts. The FAP estimate also encapsulates a wealth of information into an easily digestible form for the physics community as a whole. Claims for the detection of previously undetected physical phenomena are usually held up high statistical confidence levels of scrutiny, e.g. the Higgs boson [245] and the CMB polarisation [10]. The same will be true for direct detection of GWs: enhanced statistical confidence will be required as well as a deep understanding of the accuracy of statements made about the FAP of a candidate event.

### 6.0.1 Background of the Mock Data Challenge

In order to construct a measure of FAP it is appropriate to consider a null hypothesis. For gravitational wave detection the null hypothesis is that the data, consisting of one or more streams (time series) of estimated strain  $s(t)$  from interferometric detectors, are caused by non-astrophysical, non-gravitational-wave processes acting on and within the interferometers. In addition, we must define a test statistic such that every possible result of a given observation or experiment performed on the data is assigned a statistic value: larger statistic values indicate a higher deviation (and less probable) from expectations under the null hypothesis. Then the FAP, or *P-value*, of the result is the probability under the null hypothesis of obtaining a statistic value equal to, or larger than, the one actually obtained in the experiment.

In general the detector output data streams are the sum of terms due to non-astrophysical processes, known as background noise, and of astrophysical GW signals, called foreground. If we were able to account for all contributions to the noise using predictive and reliable physical models (for instance modelling shot noise via photon counting statistics ) we would then be able to make accurate statements on the FAP of any observation [246]. However, for real gravitational-wave detectors, in addition to terms from known and well-modelled noise processes, the data contains large numbers of non-Gaussian noise transients (“glitches”) whose sources are either unknown or not accurately modelled, and which have potentially large effects on any searches for transient GW [101]. Even given all the information available from environmental and auxiliary monitor channels at the detectors, many such noise transients cannot be predicted with sufficient accuracy to account for their effects on search algorithms. (Doing so would also require us to model the, in some cases complicated, workings of the search algorithms.) Thus, for transient searches in real detector noise it is necessary to determine the effects of noise backgrounds empirically, i.e. directly from the strain data themselves.

However, doing so has notable potential drawbacks. It is not possible to operate gravitational wave detectors so as to “turn off” the astrophysical foreground and simply measure the background; if the detector is operational then it is always subject to both sources. In addition, our knowledge of the background noise distribution is limited by random fluctuations over the finite amount of data available. This limitation applies especially in the interesting region of low probability under the noise hypothesis, corresponding to low number statistics of noise events.

Transient gravitational wave signals are expected to be well modelled by the predictions of Einstein’s general relativity [17]. They will also be observed in multiple detectors with consistent waveforms, and, crucially, with consistent arrival times. Gravitational waves propagate at the speed of light (see e.g. [2]), hence we expect the arrival times of signals to differ between detectors by  $\sim 4 \times 10^{-2}$  s or less (the light crossing time of the earth). The exact differences in arrival times are governed by the location of the source on the sky in relation to the constantly moving detectors as the Earth spins on its axis. This property acts as a very powerful test in allowing us to distinguish between true signals and detector noise or noise artefacts. It would be highly unlikely – but not impossible – that independent detector noise from 2 or more interferometers thousands of miles apart would *coincide* or conspire to appear like a signal within (say) a  $\pm 20$  ms window. The probability of such an occurrence reduces with increasing event amplitude, such that above some specific amplitude the probability of obtaining such a loud event from the noise background is small enough that the null hypothesis may be rejected in favour of an alternative hypothesis corresponding to detection of one or more astrophysical signals [149].

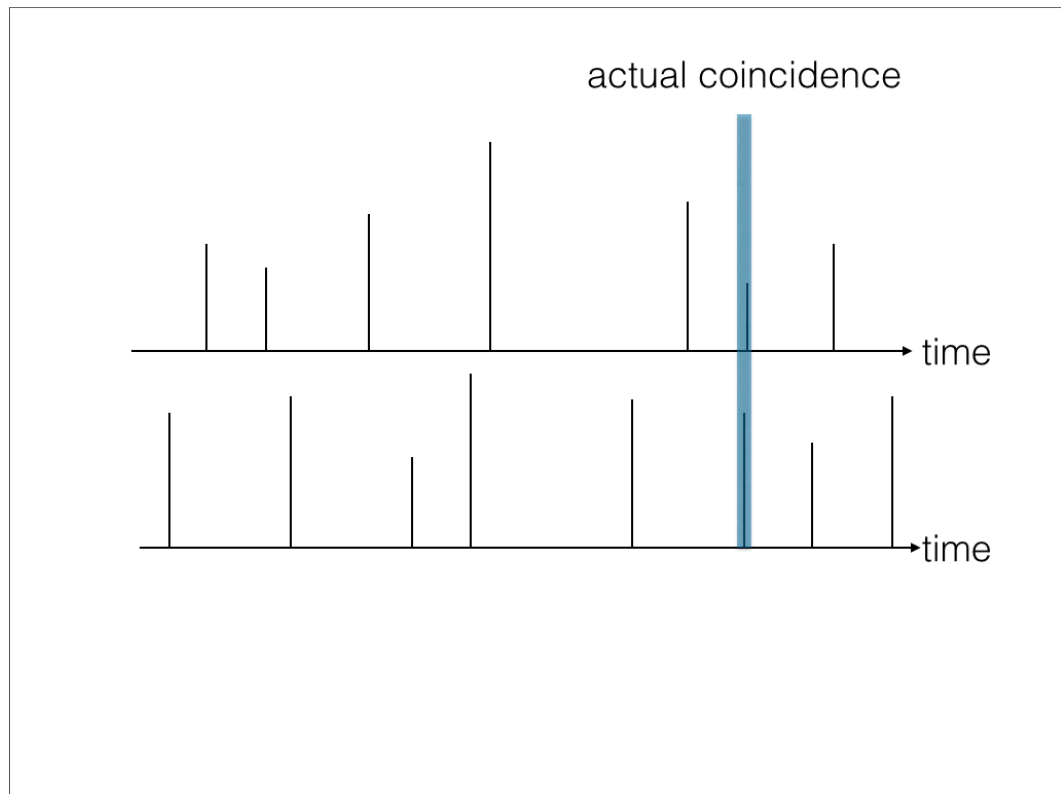
The canonical approach used in gravitational wave data analysis for estimating the background is time-shifted analysis or “time slides” [109, 247, 248]. It is an estimation method that exploits the coincidence requirement of foreground events by time-shifting the data from each detector relative to the others. Such a shift, if larger than the coincidence window as defined by the light travel time between detectors, plus possible statistical errors in the arrival times<sup>1</sup> would rule out the possibility that a foreground signal would be identified as a coincident event in a time-shifted analysis. Therefore, from a single time-shifted analysis (“time slide”) the output coincident events would represent one realisation of the background distribution. Repeated time-slide procedures using the same datasets could then be used to accumulate instances of background coincidences and thus estimate their rate and distribution.

An cartoon illustration of such time shift method is demonstrated in figure 6.1.

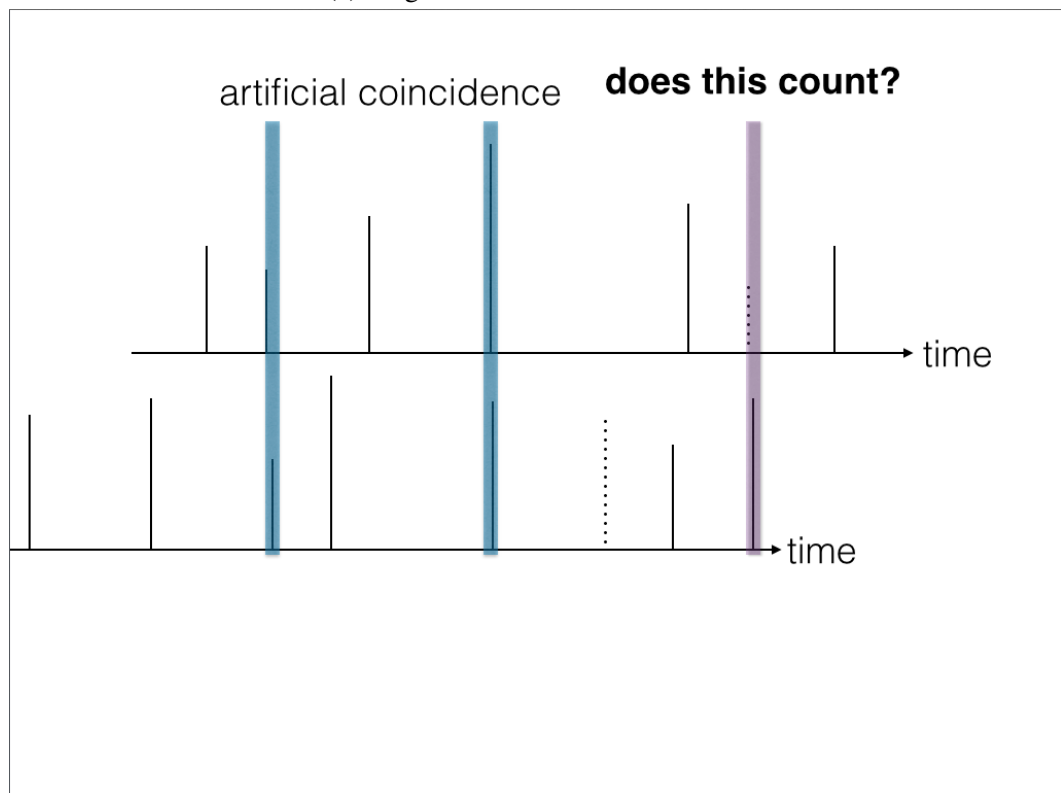
The time-slides approach has been an invaluable tool in the analysis and classification of candidate gravitational wave events in the initial gravitational wave detector era. We remind

---

<sup>1</sup>As standard, time-shifts are also large enough to allow for extended correlations between signals.



(a) Original data from two detectors



(b) Time-shifted data

Figure 6.1: Cartoon illustration of time slide method, with the vertical axes showing the SNR. The original data was observed as in figure 6.1a, if triggers from two detectors coincide within a time window, it's recognised as a coincidence. The background is estimated by shifting the time of one detector's triggers and find artificial coincidence. The controversial concentrated on the treatment of the artificial coincidence which contains the original coincidence.

the reader that there has not yet been a gravitational wave detection and hence the task of estimating event FAP has been made using data believed to satisfy the null hypothesis, or at least to contain signals which are so quiet as to be indistinguishable from noise. A notable exception occurred in 2010, when the LVC performed a self-blinded simulated signal injection [149]. A signal was injected into the global network interferometer hardware and analysed by the collaboration knowing only that there was the possibility of such an event. The impact of the blind injection for the issue of background estimation was illuminating since it highlighted potential issues with the use of time-shifted analysis in the presence of astrophysical signals (simulated or not).

Simply time-shifting detector outputs with respect to each other does not eliminate the possibility of coincident events caused by one or more foreground events in a single detector passing the coincidence test with random noise outliers in another detector. Thus, the ensemble of samples generated by time-shifted analysis may be “contaminated” by the presence of foreground events in single-detector data compared to what would result from the noise background only. The distribution of events caused by astrophysical signals is generally quite different from that of noise, in particular having a longer tail towards higher values of SNR (or other event ranking statistic used in search pipelines). Thus, depending on the rate of signals and on the realisation of the signal process obtained in any given experiment, such contamination could considerably bias the estimated background. It is expected that in this case, the estimated background level will be raised by the presence of signals; hence the FAP of coincident events seen in the search (in non-time-shifted or “zero-lag” data) may be underestimated. This leads to a conservative estimate for the FAP; in other words, the expected number of false detection claims will not increase due to the presence of signals, however some loud signals may fail to be detected due to an elevated background estimate.<sup>2</sup>

Besides the standard time-shift approach, other background estimation techniques have been developed [77, 107]. All are variants on one key concept: events that are not coincident within the physically allowed coincidence window cannot be from the same foreground source. The major variations occur in the specific implementation of the procedure. As opposed to the standard scheme, one can choose to identify all coincident events in zero-lag and flag them as potentially due to foreground events. These can then be removed before performing time-slides and hence then reducing the possibility that foreground events contaminate the background estimation. One may expect that such an approach then avoids the issue of background enhancement but it is not clear that in general this leads to any bias in subsequent FAP estimates, either conservative or otherwise.

---

<sup>2</sup>The reader may ask what a “false detection claim” means if signals are present: this refers to a situation where there are foreground events in the search, but there are also comparable or louder background noise events (with equal or higher ranking statistic values): in particular the loudest event in the search is due to background, thus a detection claim would be misleading.



In this paper we aim to rigorously test the performance of the 3 main approaches to background estimation currently used in the LVC. In each case we test each approach’s behaviour in the 2 modes of non-removal and removal of zero-lag events, the name “Hamlet Test” reflects such “To remove or not to remove” nature of the problem. The tests are undertaken in the form of a mock data challenge (MDC) in which simulated realisations of single detector triggers are generated according to analytically modelled background and foreground distributions, while such analytical expressions are kept unknown to the participants. The background distributions are chosen to model a range of potential complexity spanning and including realistic initial detector type distributions. The foreground triggers are chosen to model one of the most likely first sources, binary neutron stars, and range in their detection rate from zero up to the maximum expected for the first advanced detector science runs. Participants are asked to apply background estimation algorithms to these data datasets and report the estimated FAP of the loudest events found in each case. Since the MDC uses an analytic description of the background distributions (something we will never have access to in a realistic scenario) we can compute the “exact” FAP values semi-analytically and hence use this as our benchmark for comparison amongst other figures of merit.

In the following section give details of the MDC including descriptions of the background and foreground distributions, the data generating procedures and the details of the exact FAP calculation. In Sec. 6.2 we then describe the different background methods used in this analysis. We report the results of the challenge in Sec. 6.3, comparing and contrasting the properties of the results obtained from each of the algorithms in each of their 2 modes of operation, and finally give a summary in Sec. 6.4.

## 6.1 The mock data challenge

Data from the advanced gravitational wave detectors will be analysed using matched-filtering algorithms [16]. Although the template banks are designed to match the GW emitted during CBC inspiral stage, detector noise could mimic the signal giving rise to some level of false alarm. Only time coincident triggers among multiple detectors are considered as potential astronomical events. In order to make a confident detection claim one needs to estimate the rarity, or false alarm probability (FAP), of such an event, and only announce detection when the signal is significant enough that the possibility of it being caused by noise is extremely low.

One needs to gain knowledge regarding the SNR distribution of the triggers that are due to background noise. These triggers are the output of an incredibly complex and highly sensitive measurement instrument and their distribution is not known a priori and must be estimated empirically. Current or proposed methods described in Sec. 6.2 use a common pro-

cedure of combining non-coincident triggers from different detectors to represent possible background coincidences, and then use this distribution as an estimation of the actual distribution. However, there is a major difference among the methods, namely, whether or not to remove the actual coincident zero-lag events when estimating the background distribution. We know that the zero-lag events will likely contain background triggers, and by removing them we might be biasing our background estimate. However, if astrophysical signals are present in the zero-lag data and they are not removed then they would also act to contaminate the background estimation.

This importance of this issue was brought into attention by a blind injection signal, when hardwares are intentionally mimicing astrophysical signal and no single person knew beforehand whether it's an artificial or actual signal before the envelope was opened [149]. One such blind injection was implemented in the sixth scientific run and the two modes of estimation for the false alarm probability disagree with each other, especially on the fact whether it passes the  $5 - \sigma$  threshold. We would like to avoid such scenario for the actual detection, which motivates this work, namely to properly understand the effect of removal of zero-lag events for the FAP estimate. We have designed a MDC that seeks to resolve, amongst other issues, the question of to remove or not to remove of the zero-lag events. Simulated single detector triggers are generated artificially from semi-analytic background (and foreground) distributions and challenge participants are given lists of trigger arrival times and SNRs for each detector plus a list of coincident (zero-lag) trigger arrival times and SNRs. The CDFs of the SNR in each detector are described analytically and known only to me. With this analytic form, we are able to compute the FAP of any loudest coincidence event.

There were 14 individual simulations each consisting of  $10^5$  observational realisations and each realisation containing, on average,  $\sim 10^4$  single detector triggers. Challenge participants were asked to estimate this FAP of the loudest coincident event in each of the  $10^5$  realisations for each experiment. Between experiments the background distribution was varied as well as the astrophysical event rate which varied between zero and high values (corresponding to  $\sim 3$  detections per realisation). Analysis of the differences in the estimated and exact FAP values, enables us to quantify the influence of removal of zero-lag triggers as opposed to non-removal. It also allows us to directly compare implementation methods (since proponents of each algorithm were asked to submit both removal and non-removal based results). Finally, it allows us to quantify the expected limiting factors in the accuracy of our FAP estimates and the corresponding uncertainties.

Our aim was to explore a number of potential detector-background and foreground-level scenarios. We have therefore divided our simulations into 3 groups according to astrophysical rate, and independently into 3 groups according to background distribution complexity. Combined together there are 9 possible combinations which we have appended with an additional 4 simulations. Three of these were designed to have exactly the same background

Foreground rate	Background property			
	simple	realistic	extreme	hidden tail
low/none	1,3	10,12	2,14	7
medium	9,13	8	6	-
high	5	11	4	-

Table 6.1: The indices of all 14 experiments showing each experiment’s classification in terms of background complexity and astrophysical foreground rate. It shows a good coverage of our designed Mock Data over a variety of background complexities and foreground rates.

distributions as 3 of the original 9, and 1 simulation was included with background distributions containing an extended shallow tail. We show this division in Tab. 6.1. A simple background corresponds to a straight line relationship between single detector SNR and  $1 - \text{CDF}$  in log-space. A realistic background is achieved via fitting of our analytic CDF model to distributions of existing GW trigger data [149], and the extreme backgrounds attempt to model distributions composed of multiple populations resulting in multiple gradient changes in the distribution tails. We classify a low foreground rate as  $< 0.01$  expected coincidences per realisation, a medium rate corresponds to an expected  $0.01 - 1$  coincidence per realisation, and high rate corresponds to  $> 1$  per realisation. We do not use foreground levels that exceed  $\sim 3$  coincidences per realisation since we are interested in understanding our FAP estimation abilities for only the first advanced era detections.

### 6.1.1 Modelling the detector background

The CDF of the background SNR triggers is modelled as an exponential with argument described by a piecewise polynomial function in  $\rho - \rho_0$ , where  $\rho$  is the SNR and  $\rho_0$  is a constant. The polynomial coefficients must satisfy the constraint that the CDF remains monotonic in  $\rho$ . Additionally,  $a_0$  is subject to the constraint that the CDF should be constrained between 0–1. The CDF can be written as

$$\text{CDF}(\rho) = \begin{cases} 1 - \exp \left[ \sum_{i=0}^6 a_i (\rho - \rho_0)^i \right] & \text{if } \rho \leq \rho_{\text{SP}} \\ 1 - C_{\text{SP}} \exp [b (\rho - \rho_{\text{SP}})] & \text{if } \rho > \rho_{\text{SP}} \end{cases} \quad (6.1)$$

where we define the CDF differently in the regions below and above a switching-point  $\rho_{\text{SP}}$  in order to satisfy the constraints on the CDF model. These 2 functions are matched at the switching point such that the CDF and its first derivative with respect to  $\rho$  is continuous. Hence, a choice of  $C_{\text{SP}}$  defines the values of  $\rho_{\text{SP}}$  and  $b$ . Details regarding the background distribution parameters chosen for each simulation can be found in Chapter B; here we describe the broad properties of the distributions chosen for the MDC.

A *simple* background corresponds to a straight line relationship between single-detector SNR and  $1 - \text{CDF}$  in log-space. A *realistic* background is modelled by fitting our analytic CDF model to distributions of existing GW trigger data [149]; the *extreme* backgrounds attempt to model distributions composed of multiple populations resulting in multiple gradient changes in the distribution tails.

A background with *hidden tail* is similar to extreme in the sense that its gradient in log CDF vs.  $\rho$  varies substantially, but this variation occurs at much smaller values of  $1 - \text{CDF}$ , thus some realisations have no single event in the ‘tail’; then without considering large numbers of realisations, such a tail could be essentially hidden.

We generate the trigger time independently and randomly. Given an expected number of triggers in the  $j$ ’th detector,  $\lambda_j$ , and given that a coincident event is defined by

$$|t_1 - t_2| \leq \delta t \quad (6.2)$$

where  $t_1$  and  $t_2$  are the times associated with a trigger from the first and second detectors respectively, we can estimate the expected number of coincident events  $n$  as

$$n = \frac{2\lambda_1\lambda_2\delta t}{T}. \quad (6.3)$$

Here  $T$  is the total time of the observation and we have assumed that  $\lambda\delta t/T \ll 1$ . In order to generate a large enough number of background triggers to adequately model a realistic GW observation, we use  $\lambda_1 \sim \lambda_2 \sim 10^4$ . This choice is also motivated by the expected FAP estimation uncertainty’s dependence upon this value and the computational cost to challenge participants. We set the coincidence window  $\delta t = 50$  msec to broadly model realistic values and in order to obtain a desired  $\sim 10$  coincidences per realisation the total observation time is set to  $T = 10^6$  sec.

### 6.1.2 Modeling an astrophysical foreground

In the majority of simulations an astrophysical foreground of events is also included in each realisation. We model our astronomical signal distribution as originating from the inspiral of equal mass  $1.4 - 1.4M_\odot$  binary neutron stars governed by a uniform distribution in volume and in time. For each source the binary orientation is selected from an isotropic distribution (uniform in  $\cos \iota$ ) and the polarisation angle is uniform on the range  $[0, 2\pi)$ . With these parameters, we can predict the optimal SNR for a single detector  $\rho_{\text{opt}}$  as

$$\rho_{\text{opt}}^2 = 4 \int_{f_{\text{min}}}^{f_{\text{isco}}} \frac{|\tilde{h}(f)|^2}{S_n(f)} df \quad (6.4)$$

where the integration limits  $f_{\min}$  and  $f_{\text{isco}}$  are the lower cut-off frequency (selected as 10 Hz) and the upper cut-off frequency (selected as the innermost stable circular orbit frequency = 1570 Hz). The detector noise spectral density  $S_n(f)$  corresponds to the advanced LIGO design and the frequency domain signal, represented using the stationary phase approximation, is given by

$$\tilde{h}(f) = Q(\theta) \sqrt{\frac{5}{24}} \pi^{-2/3} \frac{\mathcal{M}^{5/6}}{d} f^{-7/6} e^{i\Psi(f)}. \quad (6.5)$$

The function  $Q(\theta)$  where  $\theta = (\alpha, \delta, \psi, \cos \iota)$  contains the antenna response of the detector. The square of the observed SNR  $\rho$  is constructed from a 2 degree-of-freedom non-central  $\chi^2$  distribution with non-centrality parameter given by  $\rho_{\text{opt}}^2$ .

We generate potential CBC events within a sphere of radius 1350 Mpc such that an optimally oriented event at the boundary has  $<0.3\%$  probability of being considered a trigger (SNR  $>5.5$ ) (the rate of such events are listed in table B.3). Each event is given a random location (uniform in volume) and orientation from which we then calculate the corresponding optimal SNR and relative detector arrival times. The observed SNR is then a random variable drawn from the non-central chi-squared distribution. For each detector, if the observed SNR is larger than 5.5, it is recorded as a single detector trigger. The arrival time in H1 is randomly selected uniformly within the observation time and the corresponding time in L1 is set by the predetermined arrival time difference (defined by the source sky position). We do not model any uncertainty in the arrival time measurements and hence, when an event is classified as a trigger in both detectors it will naturally satisfy the time coincidence criteria and would be recognised as a coincident event.

### 6.1.3 The definition of FAP for the MDC

Challenge participants were required to estimate the FAP of the “loudest” coincident event for each realisation independent of its unknown origin (either background or foreground). We combine the detection statistics from each detector, the SNRs, into a single statistic according to

$$\rho = \sqrt{\rho_1^2 + \rho_2^2} \quad (6.6)$$

We then define the FAP of the loudest combined detection statistic as the probability of obtaining a value equal to or greater than the loudest from the background distribution within a single realisation. We recall that a single realisation consists of  $\sim 10^4$  single detector triggers in 2 detectors. The FAP is a measure of how likely it is for random background noise to have generated the observed data. The louder a coincident event is, the more unlikely and the smaller the FAP will have.

Since I had access to the analytical description of the backgrounds I was able to compute the

FAP exactly by computing

$$1 - \mathcal{S}(\rho) = \int_r^{\sqrt{\rho^2 - r^2}} d\rho_1 \int_r^{\sqrt{\rho^2 - \rho_1^2}} d\rho_2 p_1(\rho_1) p_2(\rho_2) \quad (6.7a)$$

$$= \int_r^{\sqrt{\rho^2 - r^2}} d\rho_1 p_1(\rho_1) C_2 \left( \sqrt{\rho^2 - \rho_1^2} \right) \quad (6.7b)$$

Here,  $r = 5.5$  is the trigger SNR threshold and  $p_1(\rho_1)$ ,  $C_2$  is the CDF of the triggers' SNR for the second detector. and  $p_2(\rho_2)$  are the PDFs of the background distributions. We convert a two-dimensional integral over  $\rho_1$  and  $\rho_2$  to a one-dimensional integral over  $\sqrt{\rho^2 - \rho_1^2}$ . Challenge participants only had access to the triggers themselves and were not given the underlying functions from which they were drawn. Subsequent estimates of the FAP  $\mathcal{S}$  from all participants were ultimately compared to the exact values computed according to the Eq. 6.7b.

## 6.2 Background estimation algorithms

### 6.2.1 ihope: FAP via inverse false alarm rate (IFAR)

As already mentioned, searches for transient GW signals in ground-based interferometer data have extensively used time-shifted analysis to provide background noise samples and thus determine the FAP of the observed coincident events. We now describe the calculation performed in the all-sky LIGO-Virgo search pipeline for CBC and indicate how the method has been adapted for the simplified high-statistics study presented in this paper [247, 248, 107, 149, 247].

Each coincident event obtained in the search is characterised by its estimated coalescence time and binary component masses, and in addition by the values of SNR  $\rho$  and the signal-based chi-squared test  $\chi^2$  in each detector, which together are intended to achieve separation of signals from non-Gaussian noise transients [143]. The event ranking statistic used,  $\rho_c$ , is the quadrature sum of re-weighted SNRs  $\hat{\rho}_A(\rho_A, \chi_A^2)$  over participating detectors  $A$  [247, 109, 149].<sup>3</sup> Exactly the same coincidence test is performed in the time-shifted analyses as in the search, resulting in a set of values  $\{\rho_{c,b}\}$  from time-shifted events, considered as background samples.<sup>4</sup>

<sup>3</sup>In real data the search may be divided into event bins determined by the component masses and participating interferometers; these complications are irrelevant to the present study, which does not attempt to simulate them [247, 248].

<sup>4</sup>In real data an additional time clustering step is performed on the search and on each time-shifted analysis

With the search performed over a duration  $T$  of two- or more-detector data, and time-shifted analyses covering a total duration  $T_b \equiv sT$ , defining a background multiplier  $s$ , the *estimated false alarm rate (FAR)* of a candidate event having ranking statistic  $\rho_c^*$  is calculated as the observed rate of louder background events over the time-shifted analyses:

$$\text{FAR}(\rho_c^*) \equiv \frac{\sum_{\{\rho_{c,b}\}} [\rho_{c,b} > \rho_c^*]}{T_b} \equiv \frac{n_b(\rho_c^*)}{T_b}, \quad (6.8)$$

where  $[x > y]$  is equal to 1 if true or 0 if false. Note that this may be equal to zero for a high enough threshold  $\rho_c^*$ .

The test statistic used to determine FAP is inverse FAR (IFAR) defined as  $1/\text{FAR}$ ; thus, we wish to know the probability of obtaining a given value of  $n_b/T_b$ , or lower, under the null hypothesis. For small  $n_b(\rho_c^*)$  this can be shown to be

$$p(n_b(\rho_c^*)) \equiv P_0(n_b \leq n_b(\rho_c^*)) \simeq 1 - \left( \frac{s}{1+s} \right)^{n_b(\rho_c^*)+1} \quad (6.9)$$

$$\simeq \frac{1 + n_b(\rho_c^*)}{1+s}, \quad s \gg 1, \quad s \gg n_b, \quad (6.10)$$

where  $P_0(A)$  denotes the probability of  $A$  being true in the absence of signal. This result follows from taking the search events and time-shifted background events together and arguing that for each of the  $n_b + 1$  loudest such events, the probability of it occurring in the search rather than the background time is  $T/(T + T_b)$  in the absence of signals. This reasoning is based on the statistical independence of these louder events, which may not hold in general since the number of background samples obtained in time-shifted analyses may be larger than the number of single-detector triggers from which they are produced; we expect the approximation to be increasingly good as  $n_b$  becomes small [78].

The implementation is simplified considerably by considering the largest possible number of time-shifted analyses, such that a pair of single-detector triggers that are coincident in one analysis cannot be so for any other analysis; this implies that the relative time shifts are multiples of  $2\delta t$ . The maximum number of time-shifted analyses is then  $s = T/(2\delta t) - 1$ , thus  $s/(1+s) = 1 - 2\delta t/T$ . In addition the resulting time-shifted coincidences are simply the outer product of the single-detector triggers, minus those coincident in the search (“zero-lag”): so every trigger in detector 1 will be coincident with every trigger in detector 2, either in zero-lag or for some time shift. Identifying  $\rho_c^*$  with the loudest coincident search event

---

in order to reduce the number of strongly-correlated coincident events separated by short intervals ( $\lesssim 1$  s) resulting from the use of large numbers of filter templates with overlaps near unity. In this study, however, single-detector events are already uncorrelated by construction, thus such clustering is not performed.

value  $\rho_{c,\max}$  we have

$$1 + n_b(\rho_{c,\max}) = \frac{2\delta t}{T} \sum_i \sum_j [\rho_{1,i}^2 + \rho_{2,j}^2 \geq \rho_{c,\max}^2], \quad (6.11)$$

where the sums run over all single-detector triggers  $\{\rho_{1i}\}, \{\rho_{2j}\}, i = 1 \dots N_1, j = 1 \dots N_2$  and the  $+1$  on the LHS follows by including the loudest search event itself on the RHS via the  $\geq$  condition: thus the FAP estimate at the threshold  $\rho_{c,\max}$  becomes

$$p(n_b(\rho_{c,\max})) \simeq 1 - \left(1 - \frac{2\delta t}{T}\right)^{1+n_b(\rho_{c,\max})}. \quad (6.12)$$

To extend the calculation to larger  $n_b$  and generalise it to events which are not the loudest in the search we pursue a different derivation, considering the probability that a single coincident event due to background is at least as loud as a threshold  $\rho_c^*$ . This is estimated as

$$P_0(\rho_c \geq \rho_c^* | 1) \simeq \frac{\sum_{ij} [\rho_{1,i}^2 + \rho_{2,j}^2 \geq \rho_c^{*2}]}{N_1 N_2} \equiv \frac{1 + n_b(\rho_c^*)}{N_1 N_2}. \quad (6.13)$$

Here the numer 1 indicates one and only one event louder than  $\rho_c^*$ . The probability of obtaining *no* such coincident background events is a survival probability given the events obtained in the analysis; the expected number of coincident background events is  $\langle N_b \rangle = 2N_1 N_2 \delta t T^{-1}$ . Taking each coincident event as an independent trial, which will be a good approximation as long as  $N_b \ll N_1, N_2$ , the false alarm probability for  $N_b$  events is

$$p(\rho_c^* | N_b) \simeq 1 - (1 - P_0(\rho_c \geq \rho_c^* | 1))^{N_b}. \quad (6.14)$$

Since we do not know in advance how many coincident events will be present in the search, we should marginalise over this number, for which a suitable prior is the Poisson distribution

$$p(N_b | \mu) = \frac{\mu^{N_b} e^{-\mu}}{N_b!},$$

with  $\mu = \langle N_b \rangle$ . Doing so we find a notable simplification: the dependence on  $\langle N_b \rangle$  disappears and we are left with

$$p(\rho_c^*) \simeq 1 - \exp\left(-\frac{2\delta t}{T}(1 + n_b(\rho_c^*))\right), \quad (6.15)$$

which is the correct extension of Eq. (6.12) for general  $n_b$  or  $\rho_c^*$  values. For  $2\delta t/T \ll 1$  these formulae give essentially identical results at the loudest search event threshold  $\rho_c^* = \rho_{c,\max}$ . Thus, false alarms louder than  $\rho_c^*$  arising by random coincidence from our sets of single-detector triggers are approximated by a simple Poisson process with expected number



$(2\delta t/T)(1 + n_b(\rho_c^*))$ . We have verified that the  $p$ -value of Eq. (6.15) is distributed uniformly on  $(0, 1]$  for MDC data sets containing uncorrelated noise triggers. Note also that all formulae, with or without marginalisation, are consistent in the small- $p$  limit.

So far we have considered the case where all single-detector triggers are kept in constructing the background values. To implement the case of removing zero-lag coincident triggers, we simply exclude these from the sums over pairs of triggers in Eqs. (6.12) and (6.13). If the resulting  $p$ -value is equal to zero, as may happen for very loud search events  $\rho_{c,\max}$ , we regularise it by effectively adding 1 to  $n_b$  to obtain  $p = 2\delta t/T$ .

### 6.2.2 The gstlal approach

The method to estimate the FAP of coincident events based on the likelihood ratio ranking statistic described in [77] was modified for this test to use a single parameter,  $\rho_c$ . The false alarm probability for a single coincident event can be found as

$$P_0(\rho_c) = \int_{\Sigma_{\rho_c}} \prod_i p(\rho_i) d\rho_i, \quad (6.16)$$

where  $p(\rho_i) d\rho_i$  are the probability densities of getting an event in detector  $i$  with SNR  $\rho_i$ , and  $\Sigma_{\rho_c}$  is a surface of constant coincident SNR. The distributions  $p(\rho_i) d\rho_i$  are measured by histogramming the single detector SNR values either with or without the coincident events included. To get the cumulative distribution for a single event we have

$$P_0(\rho_c > \rho_c^*|1) = 1 - \int_0^{\rho_c^*} P_0(\rho_c) d\rho_c. \quad (6.17)$$

The multiple event false alarm probability is found in the same way as (6.14) where  $N_b$  is estimated from the data.

### 6.2.3 A new approach

The New Method is described in detail in [107]. Here we provide a brief synopsis.

To estimate the false alarm probability of foreground triggers, we first find the probability of getting a trigger from the background distribution with combined SNR  $\geq \hat{\rho}$  in a single draw. When not removing foreground triggers from the background estimate, this is:

$$\mathcal{F}(\hat{\rho}) = \frac{n(\hat{\rho})}{N_1 N_2 - k}. \quad (6.18)$$

Here,  $\hat{\rho}$  is the quadrature sum of the single-detector statistics,  $N_i$  is the total number of

triggers in the  $i$ th detector,  $k$  is the number of foreground triggers, and  $n(\hat{\rho})$  are the number of background triggers with a combined SNR  $\geq \hat{\rho}$ .

Both background and foreground triggers are constructed by finding every possible combination of triggers in detector 1 and detector 2. Background triggers are then any coincidence such that  $\Delta t = |t_1 - t_2| > w$ , where  $w$  is the coincidence window, while foreground triggers are those with  $\Delta t \leq w$ . These can be found by adding the matrices  $Z = X + Y$ , where  $X_{ij} = \rho_1^2[t_i] \forall j$  and  $Y_{ij} = \rho_2^2[t_j] \forall i$ . The elements of  $Z$  are thus the  $\hat{\rho}^2$  of all possible combination of triggers; those  $Z_{ij}$  with  $|t[i] - t[j]| \leq w$  are foreground; the rest background.

When removing foreground triggers from the background, the single detector triggers that form the foreground are removed from the  $X$  and  $Y$  matrices prior to finding  $Z$ . This changes the denominator in Eq. (6.18) to  $(N_1 - k)(N_2 - k)$ . However, if  $N_1, N_2 \gg k$ , then the denominator is approximately  $N_1 N_2$  in either case; we use this approximation in the following.

Since Eq. (6.18) is a measured quantity, it has some uncertainty  $\delta\mathcal{F}$ . This is given by:

$$\pm \delta\mathcal{F} \left( \hat{\rho} = \sqrt{\rho_1^2 + \rho_2^2} \right) = \mathcal{F} \sqrt{\sum_{i=1,2} \left( \frac{\pm \delta F_i(\rho_i)}{F_i(\rho_i)} \right)^2}, \quad (6.19)$$

where  $F_i(\rho_i)$  is the estimated survival function in the  $i$ th detector, given by:

$$F_i(\rho_i) = \frac{n_i(\rho_i)}{N_i}. \quad (6.20)$$

Here,  $n_i(\rho_i)$  is the number of triggers in the  $i$ th detector with SNR  $\geq \rho_i$ . We estimate  $\delta F_i$  by finding the range of  $F_i$  for which  $n_i$  varies by no more than one standard deviation. Using the Binomial distribution this is:

$$\max_{\min} F_i = \frac{N_i(2n_i + 1) \pm \sqrt{4N_i n_i(N_i - n_i) + N_i^2}}{2N_i(N_i + 1)}. \quad (6.21)$$

The error is thus:

$$\pm \delta F_i = \mp F_i \pm \max_{\min} F_i. \quad (6.22)$$

This error estimate can be asymmetric about  $F_i$ ; to propagate to  $\delta\mathcal{F}$ , we use  $+(-)\delta F_1$  and  $+(-)\delta F_2$  to find  $+(-)\delta\mathcal{F}$ .

Equation (6.18) estimates the probability of getting a trigger with combined SNR  $\hat{\rho}$  in a *single* draw from the background distribution. If we do  $k$  draws, the probability of getting one or more events from the background with combined SNR  $\geq \hat{\rho}$  is:

$$\overline{\mathcal{F}}(\hat{\rho}) = 1 - (1 - \mathcal{F}(\hat{\rho}))^k, \quad (6.23)$$

with error:

$$\pm \delta \overline{\mathcal{F}}(\hat{\rho}) = k(1 - \mathcal{F})^{k-1}(\pm \delta \mathcal{F}). \quad (6.24)$$

Thus, if we have two detectors with  $N_1$  and  $N_2$  triggers,  $k$  of which form foreground, or correlated, coincidences, then we can estimate the probability (and the uncertainty in our estimate of the probability) that each trigger was drawn from the same distribution as background, or uncorrelated, coincidences using Eqs. (6.18)–(6.24). The smaller this probability is for a foreground coincidence, the less likely it is that that coincidence was caused from uncorrelated sources. Since gravitational waves are expected to be the only correlated source across detectors, we use this probability as an estimate for the false alarm probability.

As this study is concerned with just the loudest foreground events, it is useful to evaluate the smallest false alarm probability that can be estimated using this method, and its uncertainty. From Eq. (6.18), the smallest single-draw  $\mathcal{F}$  that can be estimated is  $(N_1 N_2)^{-1}$ . By definition, this occurs at the largest combined background SNR,  $\hat{\rho}^\dagger$ . If the combined SNR of the loudest foreground event is not  $> \hat{\rho}^\dagger$ , then  $\hat{\rho}^\dagger$  must be formed from the largest SNRs in each detector, so that  $n_i = 1$ . Assuming  $N_1, N_2 \gg 1$ , then from Eqs. (6.23) and (6.19) we find:

$$\min \overline{\mathcal{F}} \pm \delta \overline{\mathcal{F}} \underset{N_{1,2} \gg 1}{\approx} \frac{k}{N_1 N_2} \left[ 1 \pm \left\{ \begin{array}{c} 2.3 \\ 0.87 \end{array} \right\} \right]. \quad (6.25)$$

If the combined SNR of the loudest foreground is  $> \hat{\rho}^\dagger$ , then we cannot measure its false alarm probability. In this case, we use Eq. (6.25) to place an upper limit on  $\overline{\mathcal{F}}$ .

Determining the  $\overline{\mathcal{F}}$  for every foreground trigger can require storing and counting a large number of background triggers. To save computational time and storage requirements, we reduce the number of background triggers that have  $\mathcal{F} > \text{some fiducial } \mathcal{F}_0$  by a factor of  $\mathcal{F}/\mathcal{F}_0$  for each order of magnitude increase in  $\mathcal{F}$ . We then apply a weight of  $\mathcal{F}/\mathcal{F}_0$  to the remaining background triggers when finding  $\mathcal{F}$  for the foreground. For this study,  $\mathcal{F}_0$  was chosen to be  $10^{-5}$ . Thus, between  $\mathcal{F} = 10^{-4}$  and  $10^{-5}$ , 1 out of every 10 background triggers was kept, with a weight of 10 applied to the remaining. Likewise, between  $\mathcal{F} = 10^{-3}$  and  $10^{-2}$ , 1 out of every 100 background triggers was kept, with a weight of 100 applied to the remaining; etc. This substantially reduces the number of background triggers that need to be counted and stored; e.g., for  $N_1 N_2 = 10^8$ , only  $\sim 5000$  background triggers are needed, a savings of about 5 orders of magnitude. The trade-off is that our accuracy in measuring the false alarm probability is degraded for triggers with  $\mathcal{F} > \mathcal{F}_0$ . This is assumed to be acceptable in a real analysis, since triggers with larger  $\mathcal{F}$  are, by definition, less significant.<sup>5</sup>

<sup>5</sup>In retrospect, this background degradation was not really necessary for this study, since we were only interested in the false alarm probability of the loudest foreground event in each realisation. However, we wished to keep the analysis method as similar as possible to what would be done in a real analysis.

## 6.3 Results

In order to achieve our aims it is necessary for us to examine multiple properties of the submitted challenge results. In the following sections I first investigate the accuracy of the FAP estimates by direct comparison with the exactly known true values for each realisation in each simulation. I then examine the self-consistency of each set of results for each simulation. Since the reported quantity is an estimate of a probability of an event in the absence of signal, it is expected that the rate of occurrence of FAP values below a certain threshold should be equal to their value. I then construct Receiver Operating Characteristic (ROC) plots for each experiment as a way to compare estimates in terms of their detection efficiency at fixed false positive rates. Finally I address the general issue of FAP estimate precision and attempt to extrapolate our findings to predict the likely uncertainties on our confidence in GW detection.

The challenge was attempted by 3 different teams, each of which used similar but independent algorithms (see Sec. 6.2). Each team was asked to operate their algorithms in 2 modes, one in which zero-lag triggers were included in the background estimation and the other in which they were removed. For each realisation of each simulation this gives us 6 FAP estimates to compare. In all subsequent results sections we include only plots from simulations that highlight the main features of our comparative analysis. All other plots can be found in chapter A.

### 6.3.1 Direct comparison with exact FAP

In this section, we show the direct comparison of FAP estimation with the exact FAP. Estimation methods aim to have both good accuracy and good precision, i.e., the spread should be small and the estimation should concentrate around the exact value. The estimated values could be influenced by a number of factors including random fluctuation from being derived triggers, themselves derived from actual noisy data. There may also be hidden structures within the background distribution that were not promised to be sampled by the triggers, and there may be contamination from a population of foreground triggers. Where possible we attempt to link the performance of the algorithms with these factors, and to understand and quantify their influences.

As both *gstlal* and *ihope* method used a hard cutoff in the estimation of FAP, there are visible flat lower boundaries for *gstlal* and *ihope* in figure 6.3, 6.5 and 6.6.

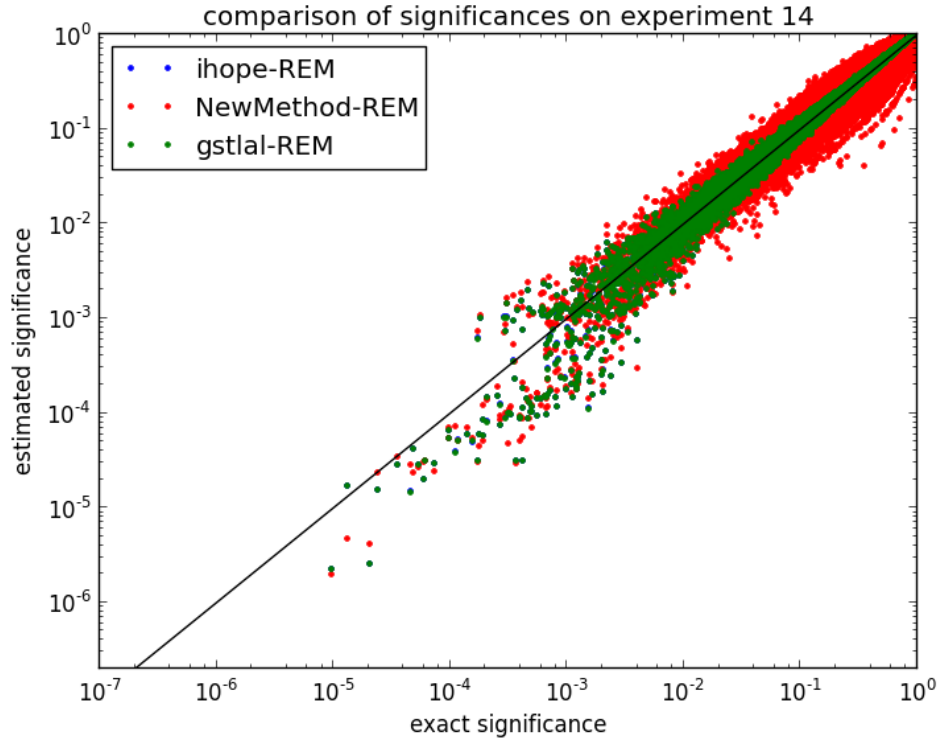
### Low foreground rate

For the experiments with low foreground rate or even no foreground signals, the triggers are only slightly contaminated by foreground signals, so the estimation of the background should be correspondingly unaffected. In cases where there are no foreground triggers present, even the extreme backgrounds, e.g., experiment 14, shown in Fig. 6.2, don't appear to affect the estimation and the spread is relatively small around the diagonal line. Notice that in these direct comparison figures, a perfect estimation will give exactly the same FAP as the true value, so it will lie on the diagonal line. If an algorithm provides an overestimate by assigning a smaller FAP, it will fall below the diagonal line, and underestimation will lie above the diagonal line. In Fig. 6.2 we see that estimation from *gstlal* is very close to result from *ihope*. For other experiments, their results do show some small discrepancy of different lower boundaries to the estimation. For all submitted results from the New Method, the estimation method used was designed such that only for rare events (those with low FAP values) were the results computed to the highest accuracy. This is motivated by the fact that the astrophysical events that we are ultimately interested in will necessarily have small FAP values and by the computational load of the challenge itself. This would lead to the enhanced scatter for larger FAP visible in figure 6.2, 6.3, 6.5 and 6.6. Notice that in the removal case for all algorithms, there is a tendency to underestimate the FAP, This is non-conservative in the sense that underestimates indicate that events are more rare than in reality.

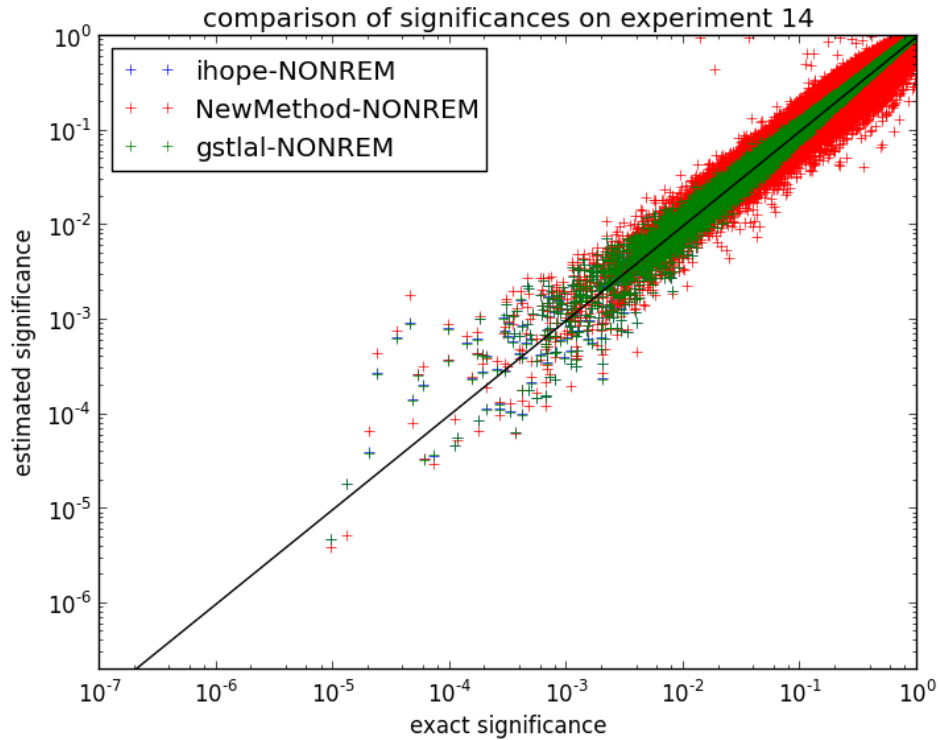
### Medium foreground rate

We class experiments as having a medium foreground rate as those with an average of  $\sim$ half the realisations containing a foreground coincidence. In the cases where foreground triggers exist it is expected that if the estimation of FAP is biased by this contamination then it would likely be biased toward overestimation. Foreground triggers are drawn from a relatively shallow distribution in SNR and are likely to be loud if present. If they are included in the background estimation it would incorrectly enhance the number of loud background events leading to an overestimate of FAP. This kind of bias would be considered to be conservative since it would underestimate the rarity of possible astrophysical events.

We use results from experiment 9 and 6, shown in Figs. 6.5 and 6.6, as examples of a medium foreground rate. In the first case we see greater variation in the FAP estimates specifically in the low FAP region in comparison to low foreground rate cases. This is clearly caused by the presence of foreground events since nearly all points below  $10^{-5}$  on the x-axis are due to foreground signals. We again see general agreement between algorithms (with the exception of the New Method at high FAP) but note that there is now evidence of discrepancies at the lowest values of estimated FAP. This is mostly due to the different choices of lowest estimable value between algorithms and is independent of the MDC dataset. There is now also



(a) direct comparison with removal



(b) direct comparison with non-removal

Figure 6.2: Direct comparisons of FAP estimates with the true FAP for experiment 14 (containing no foreground signals). Despite the background distribution being classed as extreme the estimation is concentrated on the diagonal in both removal and non-removal cases. Notice though that for removal for all algorithms the points at low true FAP values ( $<10^{-3}$ ) tend to be under the diagonal line resulting in non-conservative FAP estimates. In both plots the majority of blue points are masked by the green points since these methods provide closely matching results.

stronger indication that removal shifts estimates to lower values compared to non-removal. We remind the reader that these plots are plotted in logarithmic scales. Further investigation of this possible bias is described in Sec. 6.3.4.

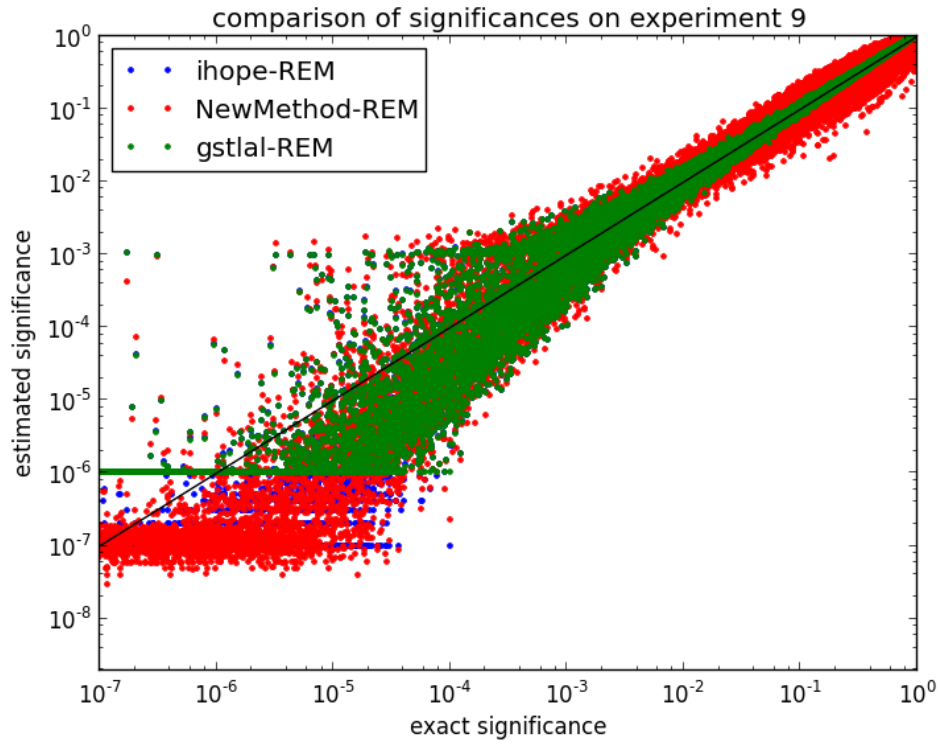
The experiment shown in Fig. 6.5 contained a background distribution that featured a shallow platform (which can be seen in Fig. 6.4a) and was classed as an extreme background. Here we see similar behaviour as seen for experiment 9 but with greater variation in estimated values that span  $\sim 4$  orders of magnitude for all true FAPs below  $\sim 10^{-3}$ . In some of the realisations from this experiment samples are rarely or never drawn from around the platform feature in the background which appears where the single detector CDF  $\sim 10^{-1}$ . In these cases where in a single realisation there is no evidence in the triggers themselves regarding any shallow features then algorithms can be misled towards large overestimation, while in others realisations the contamination of foreground would lead to strong underestimation.

### High foreground rate

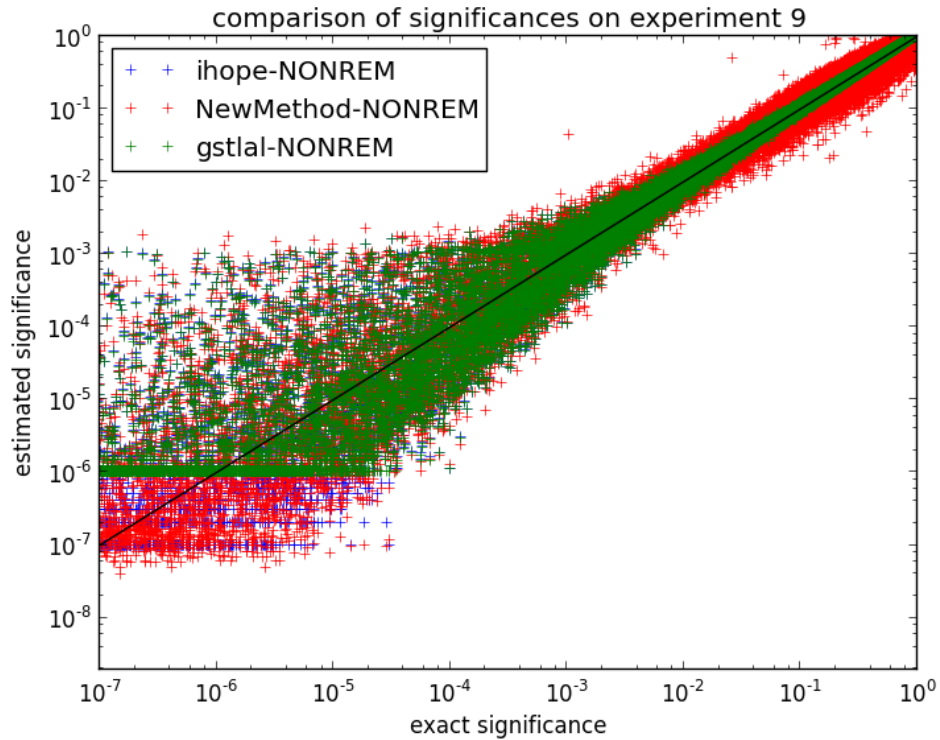
We class high foreground rate experiments as those containing realisations in which there is likely to be  $>1$  foreground event. Results from experiment 11 are shown in Fig. 6.6 where as is generally the case, there is good agreement between algorithms but clear variation between removal and non-removal modes. For removal, the presence of many contaminating foreground signals shifts estimates to higher values. This shift reaches the point where, in this experiment, the bulk of the distribution now appears consistent with the expected values with relatively narrow spread. We remind the reader again that this log-space representation can be misleading. For the non-removal case the contamination effect is greater and so is the corresponding shift to higher values of FAP (a conservative shift). In fact, in this case underestimates of the FAP are extremely rare. For all algorithms and for both removal and non-removal, as the foreground rate increases, a noticeable horizontal feature appears in these comparison plots. We discuss this in the next section.

### Horizontal bar feature

We notice that in all high foreground rate scenarios there are visible horizontal features appearing at  $\sim 10^{-3}$  in estimated FAP, which is also marginally visible in medium rate situations. The process of FAP estimation for the loudest coincident event is based on collecting the fraction of all possible unphysical coincidences which are louder. Consider a single trigger in 1 detector that is extremely loud and not found in coincidence in zero-lag. It is now highly likely that triggers from the other detector will combine into an artificial coincidence being louder than the original loudest (the details of this process would be specific to each algorithm). Assuming a method that makes use of all possible unphysical trigger combinations



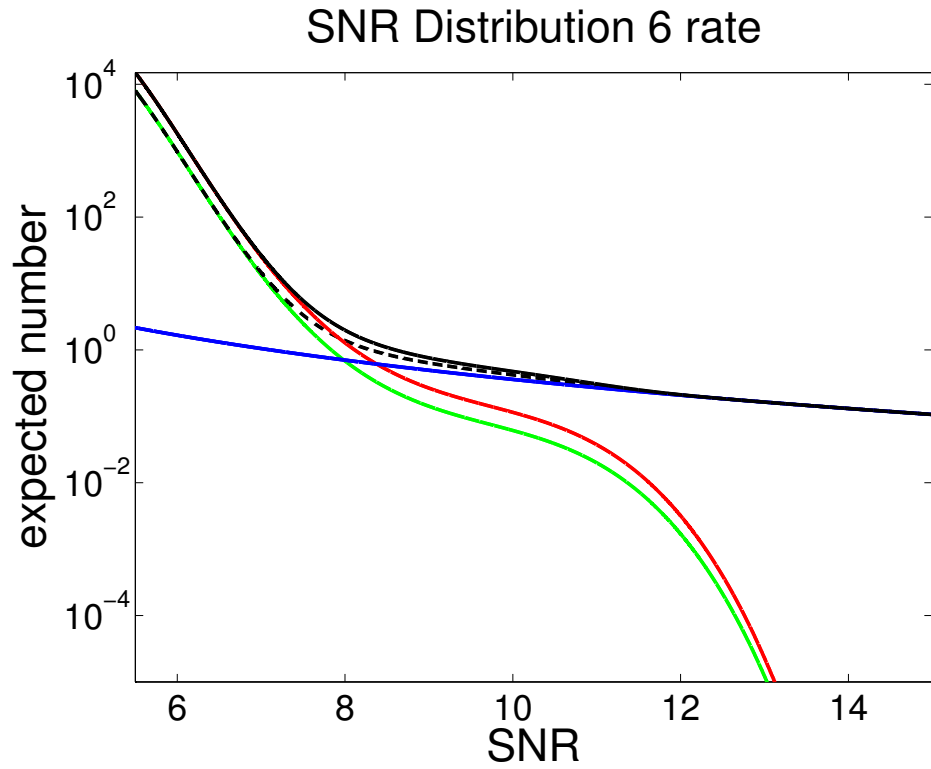
(a) direct comparison with removal



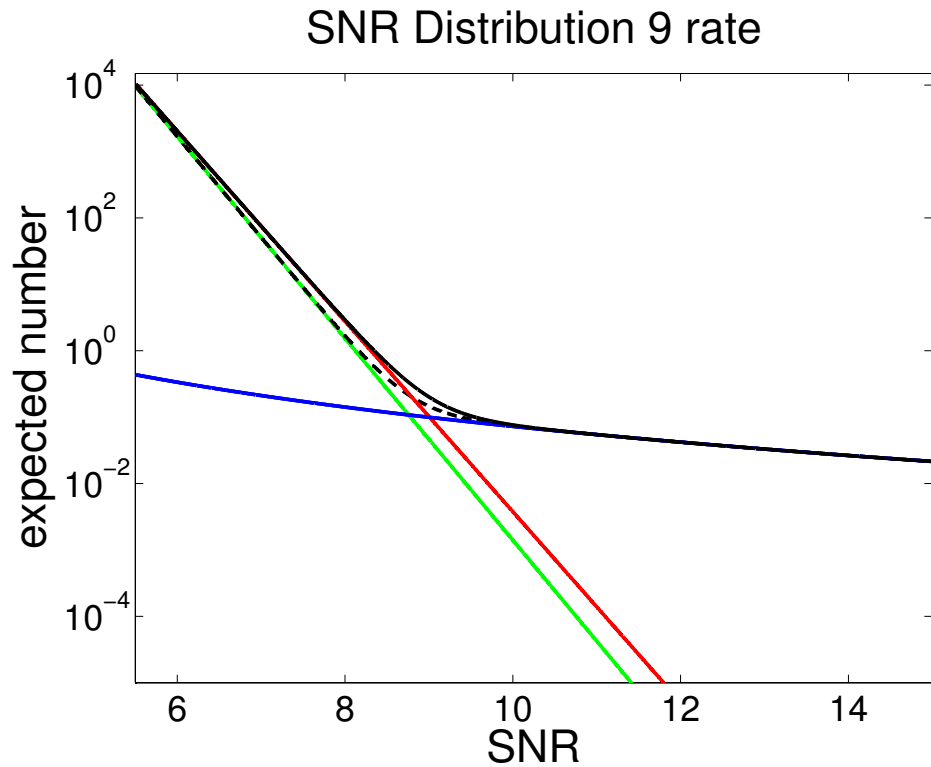
(b) direct comparison with non-removal

Figure 6.3: Direct comparisons of FAP for experiment 9. The medium level foreground rate in this case leads to some realisations being contaminated by foreground signals. Together with the cases where no foreground signal is present while others are not, thus causing a larger spread vertical spread.



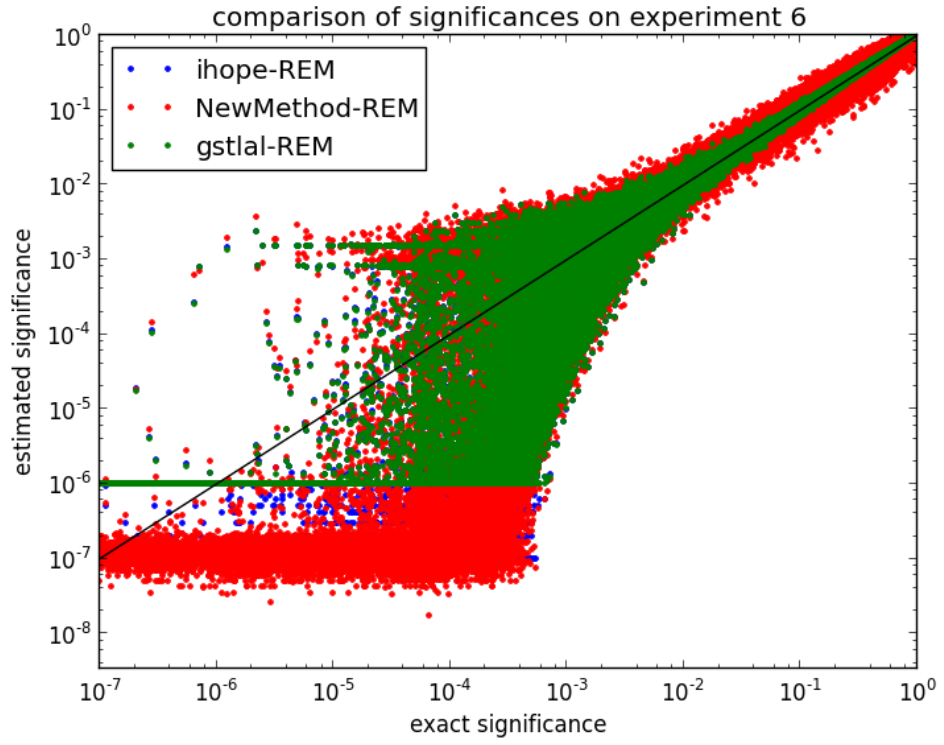


(a) Reverse CDF of SNR for experiment 6.

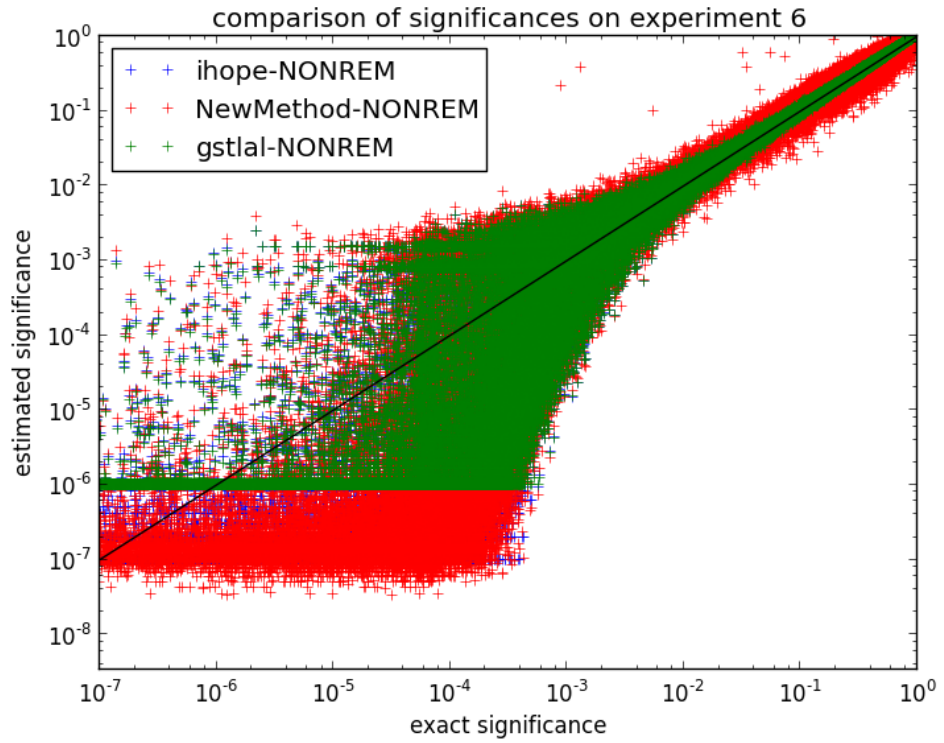


(b) Reverse CDF of SNR for experiment 9.

Figure 6.4: Reverse CDF of the triggers' SNR for experiment 6 and 9. The red and green curves represent the two individual detectors, while the blue curve represents the astronomical signals. The black lines represent the combined distribution of both background and foreground triggers, with the straight line and the dashed line corresponding to two detectors.

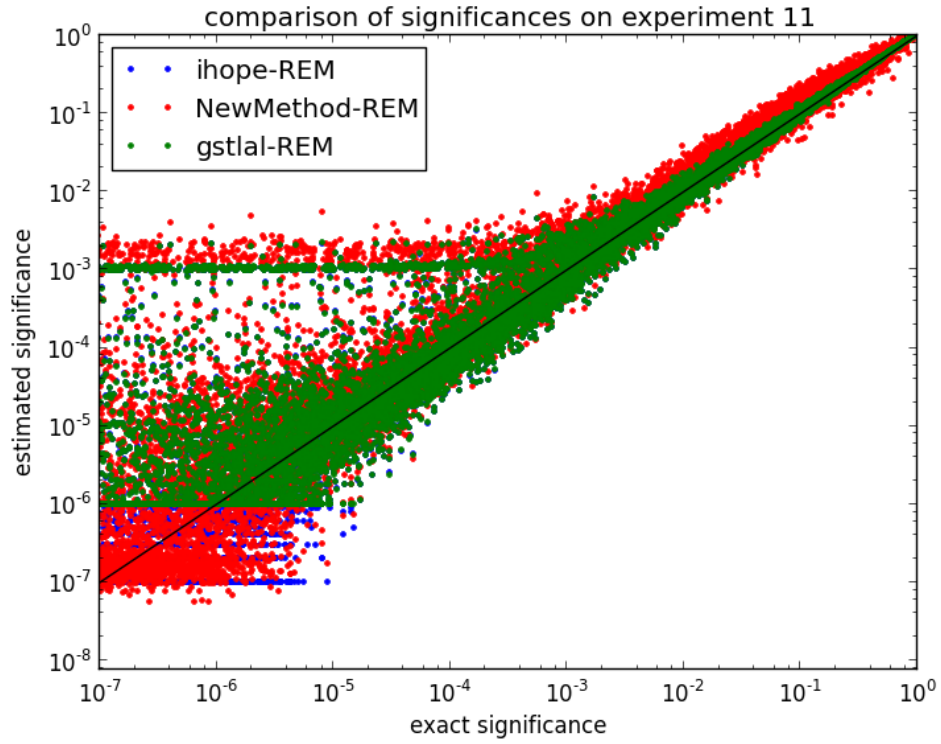


(a) direct comparison with removal

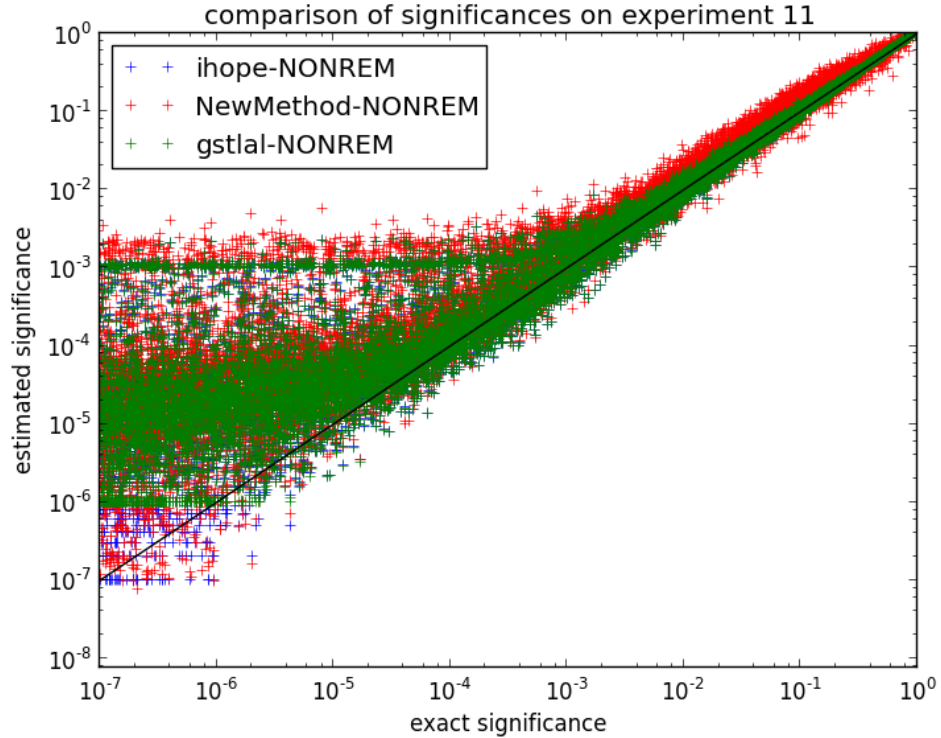


(b) direct comparison with non-removal

Figure 6.5: Direct comparisons of FAP for experiment 6. The presence of a platform feature in the tail of the distribution makes the spread in estimation wider than for experiment 9.



(a) direct comparison with removal



(b) direct comparison with removal

Figure 6.6: Direct comparisons of FAP for experiment 11. The high foreground rate in this case causes general shifts towards larger estimates for the FAP. The horizontal bar feature visible in both plots is also most prominent in this high rate case.

between detectors, this corresponds to  $\sim 10^4$  louder artificial coincidences out of a possible  $\sim 10^8$  in total. Considering an expected  $\sim 10$  zero-lag coincidences this gives an estimated FAP of these asymmetric events as  $\sim 10^{-3}$ , causing such horizontal bar feature.

In experiment 11 (Fig. 6.6), there are  $\sim 650$  such realisations. For  $\sim 500$  of them, the cause is a single astrophysical foreground trigger appearing as an extremely loud event in one detector, while for the other detector, the combination of antenna pattern and non-central  $\chi^2$  random fluctuations results in a sub-threshold SNR and is hence not recorded as a trigger. The remaining  $\sim 150$  events also have very loud SNRs in one detector, but in these cases the counterpart in the second detector appears only as a relatively weak trigger. Therefore, in performing unphysical coincidences for background estimation, almost all combinations involving the loud single detector event are louder than the actual loudest recorded zero-lag coincident event. For those foreground events that do appear asymmetrically in both detectors, a removal method could mitigate this effect; while for the  $\sim 500$  events that contain a single loud foreground event, this effect will always happen even for removal methods.

### 6.3.2 Self consistency tests: $p$ - $p$ plots

In Fig. 6.7-6.10 we show the relationship between the estimated FAP values and their cumulative frequency of occurrence. When the events are drawn from the background distribution that the FAP values are estimating then we expect the curves to trace the diagonal. The figure shows results for the 4 experiments (1,3,12 and 14) for which there were only background triggers. As we probe lower FAP values (i.e., rarer events) we begin to see measurement noise due to the finite number of rare events. However, we see a marked difference between the removal and non-removal approaches and no discernible differences between methods. In all cases the non-removal scheme stays consistent with the diagonal within the expected noise fluctuations due to the finite number of samples. The removal results however, always systematically underestimate the FAP for the rarest events and deviation from the expected behaviour occurs for all FAP values below  $\sim 10^{-3}$ .

Experiments 1 and 3 were both designed to have simple background distributions and therefore the log of their CDF tails are linear in SNR, each experiment with a different slope and with each detector having the same distribution. Experiment 14 was designed to have an extreme background distribution with multiple CDF features. The behaviour of the  $p$ - $p$  plots in all 3 of these cases is very similar with removal methods deviating from the diagonal for FAPs  $< 10^{-3}$  and in this region being  $\sim 1$ – $2$  standard deviations from the diagonal. At true FAP values of  $10^{-4}$  the removal methods tend to contain  $\sim 3$  times as many events estimated to be at or below this value. The non-removal methods remain consistent throughout within their  $1$ - $\sigma$  uncertainties. For experiment 12, containing a realistic background distribution, deviation from the diagonal occurs at approximately the same point but the deviation at a

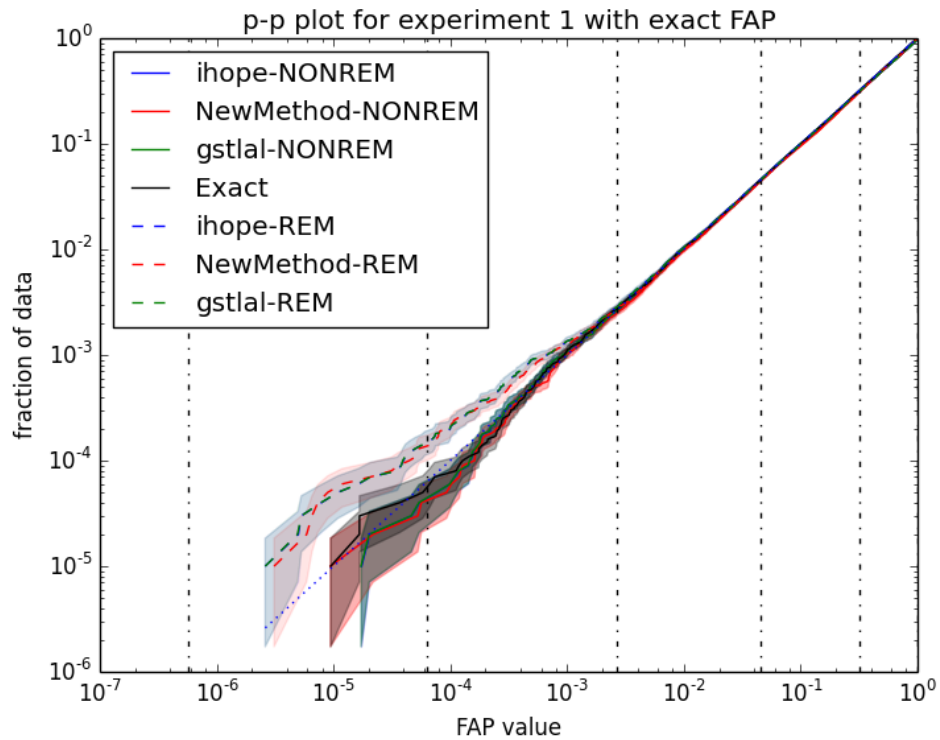


Figure 6.7: Plots of estimated FAP value versus the fraction of events with that value or below (known as a  $p$ - $p$  plot) for experiment 1. If the estimate of FAP is accurate we would expect the value to be representative of its frequency of occurrence and hence the diagonal line indicates a perfect FAP estimate. We show results for the experiments where the triggers were generated from background only. The solid lines are the results obtained for our 3 algorithms in non-removal mode whilst the dashed lines are for the removal mode of operation. Vertical dashed lines indicate the FAP associated with integer multiples of Gaussian standard deviations, i.e., the equivalent of  $n$ - $\sigma$  confidence.

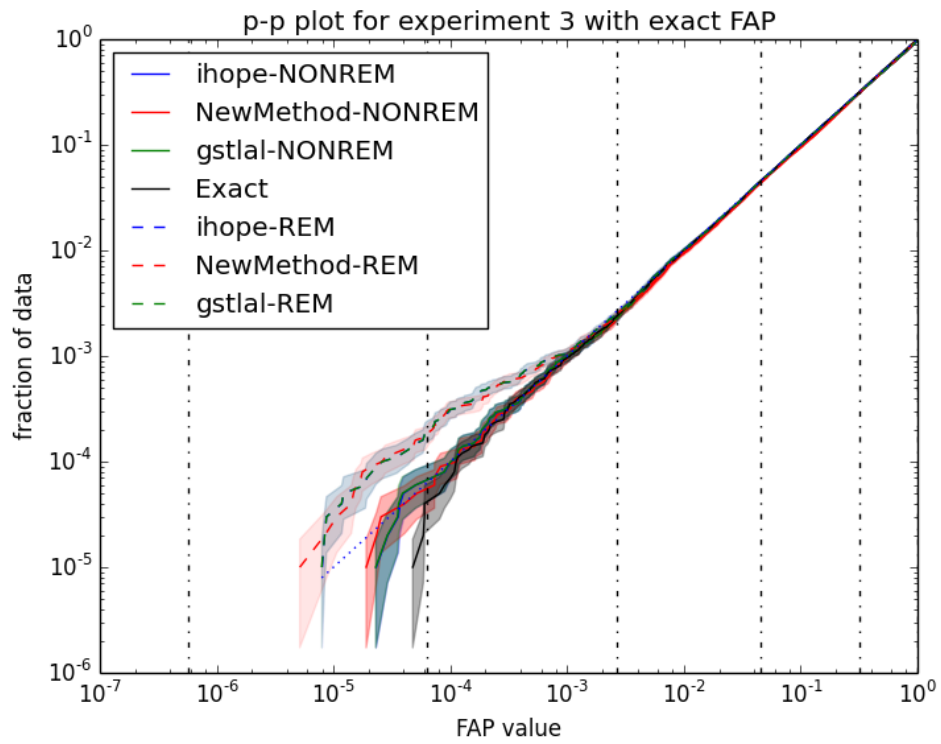


Figure 6.8:  $p$ - $p$  plot for experiment 3. Figure plotted with the same method as figure 6.7.

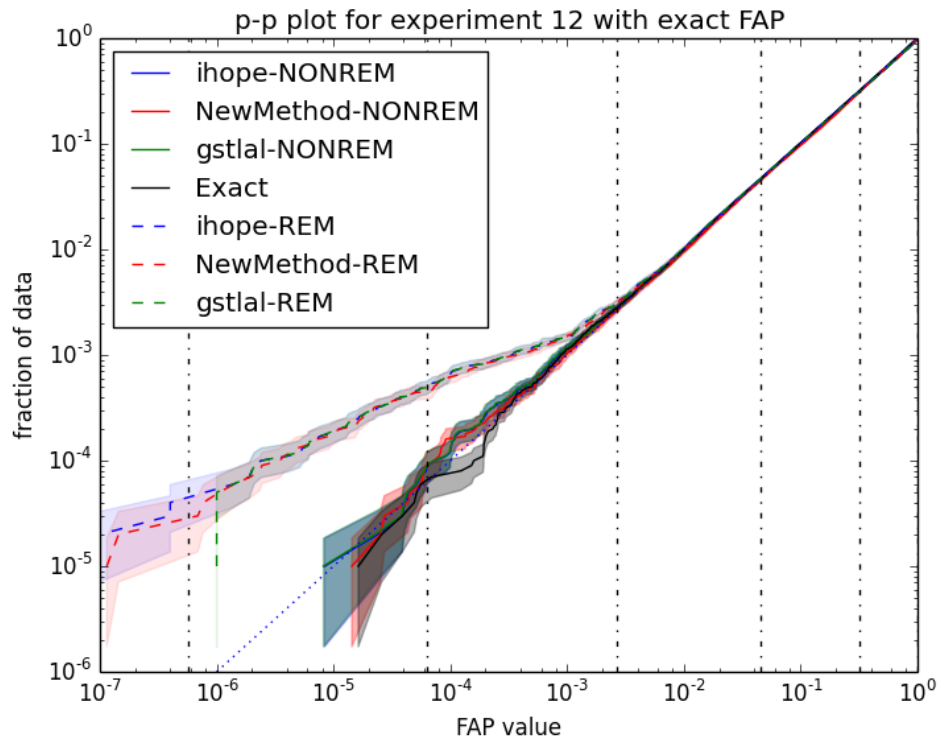


Figure 6.9:  $p$ - $p$  plot for experiment 12. Figure plotted with the same method as figure 6.7.

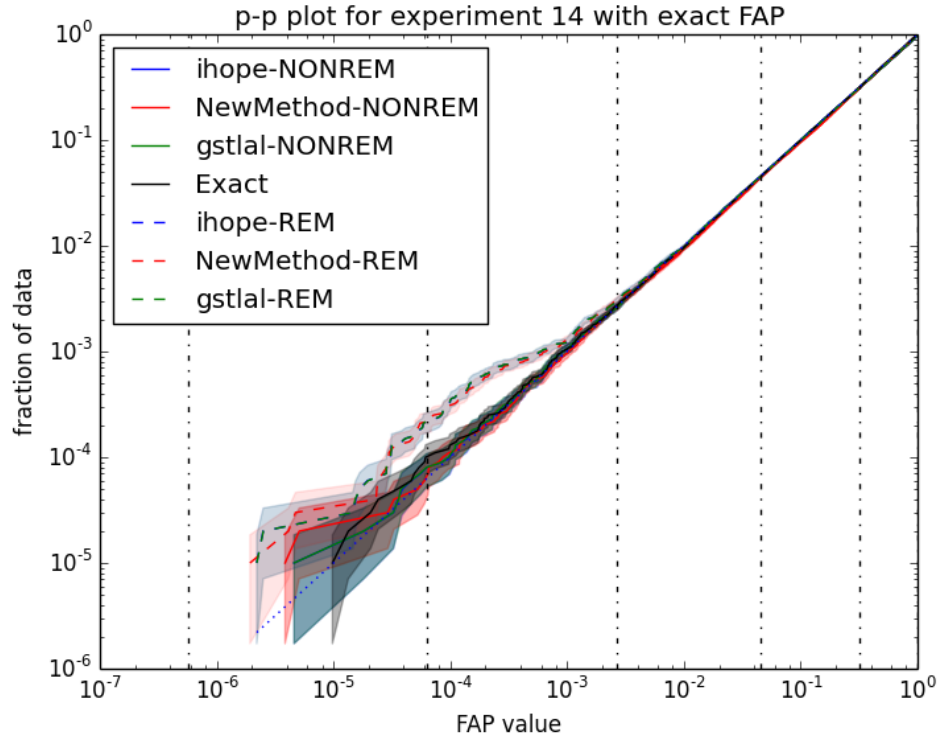


Figure 6.10:  $p$ - $p$  plot for experiment 14. Figure plotted with the same method as figure 6.7.

true FAP of  $10^{-4}$  implies that there are  $\sim 7$  times the number of estimated values at or below this level. In addition, the deviation in this case is in terms of statistical uncertainty on the curves is equivalent to  $\sim 10$  standard deviations and in all experiments cannot therefore be explained by counting fluctuations.

The deviations seen for the removal case do not make any statements on the individual point estimates of FAP for specific realisations. Hence it does not directly indicate any bias in those estimates. It does however indicate that for rare events in a background only dataset, using a removal method gives a greater than  $\mathcal{S}$  chance of obtaining an event of estimated FAP  $\mathcal{S}$ .

### 6.3.3 Receiver Operating Characteristic (ROC) analyses

The FAP quantity is designed to convey a measure of rarity of observed events but in this section we treat the FAP as a test statistic. Doing this allows us to use Receiver Operating Characteristic (ROC) plots to compare the ability of distinguishing signal from noise for each method. The ROC plot is a method to compare false negative rates given the same false positive rate criterion. In practice this involves taking one of our experiments containing  $10^5$  realisations and as a function of a variable threshold on our test statistic (the FAP) computing the following. The false positive rate (FPR) is the fraction of loudest events that were due

to background that were below the threshold. In our case having lower FAP is indicative of rare events and therefore of signals and therefore background below threshold is classed as a false positive. The true positive rate (TPR) is computed as the fraction of loudest events that were due to the foreground that have estimated FAPs below the threshold. For each choice of threshold a point can be placed in the FPR-TPR plane creating an ROC curve for a given test-statistic. Better performing statistics produce curves that appear higher in the plot such that at a given FPR the TPR is greater than for other statistics. A perfect method would recover 100% of the signals, while incurring no false alarms, corresponding to a ROC curve that passes through the left-upper corner. A totally ignorant classifier would assign values to its statistic randomly and the FPR and TPR would be identical giving a diagonal ROC curve.

In general, as can be seen in our ROC plots (figure 6.11-6.13), as the foreground rate goes up, the more events are available to compute the TPR, thus making the vertical uncertainties smaller. Conversely, the more background events are available, the smaller the horizontal uncertainties. Error-bars are computed using a binomial likelihood function in both the horizontal and vertical directions. Also, in the following subsections we focus on the experiments where there are clear discrepancies. The cases where there is agreement between methods are shown in chapter A.3. Finally we stress that ROC curves allow us to assess the ability of a test-statistic to distinguish between different distributions. It makes no direct statement on the accuracy or precision of FAP estimation.

### Low foreground rate

There are 3 experiments, 2, 7 and 10, that have non-zero low foreground rates. We show the ROC curve for experiment 2 in Fig. 6.11 which shows a number of interesting features. Firstly we note that there is general agreement between algorithms and deviations are only visible between removal and non-removal modes of operation. At a FPR of  $\sim 10^{-3}$  and below, the non-removal schemes appear to achieve higher TPRs when accounting for their respective uncertainties by  $\sim 10\%$ . This indicates that in this low-rate case where  $\approx 1$  in a 1000 loudest events were actual foreground signals the non-removal approach is more efficient at detecting signals at low FPRs. We can actually identify all experiments that show such deviations, and they all have tail feature or obvious asymmetry between two detector's background distribution, combined with a low to medium foreground rate. For Experiment 10, shown in Fig. A.27 the foreground rate was an order of magnitude lower and hence the low number of foreground events results in the uncertainty on the TPR being too great to infer any differences between approaches. Experiment 7 contained a hidden extended shallow tail in the background distribution included to mimic a foreground distribution. This did not prove to be problematic in terms of this ROC curve analysis (shown in Fig. A.25) and all algorithms and removal and non-removal methods were in good agreement.



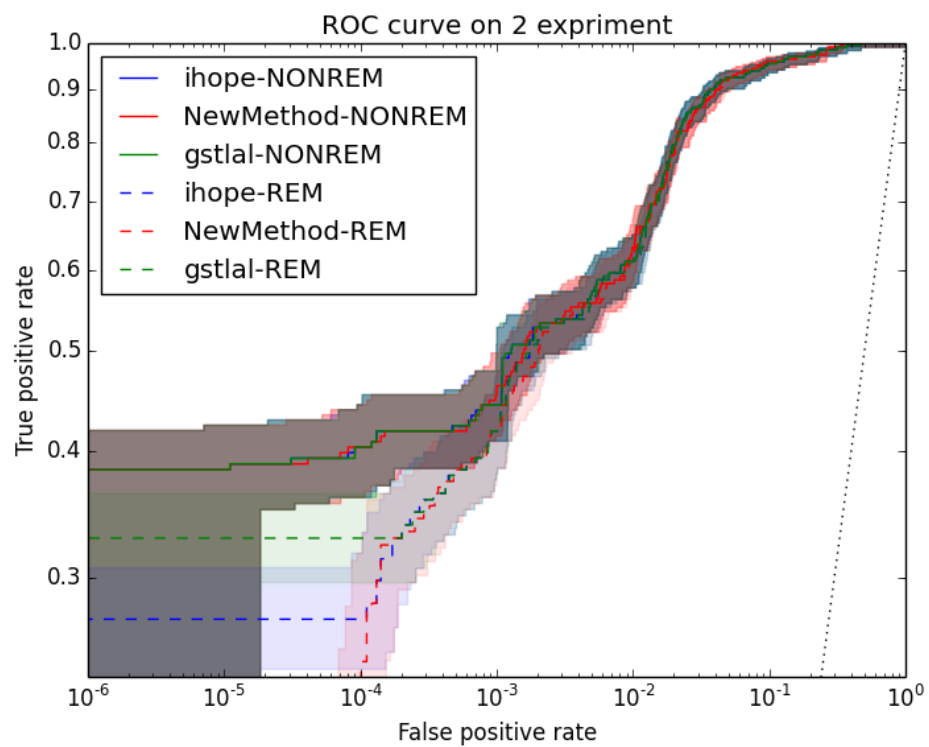


Figure 6.11: ROC plot for experiment 2. The error bars in both horizontal and vertical direction are calculated with a binomial likelihood under a 68.3% credible interval. For each point, the outer error-bar of horizontal and vertical will be picked up to be shown. The solid lines correspond to non-removals while dashed lines removal.

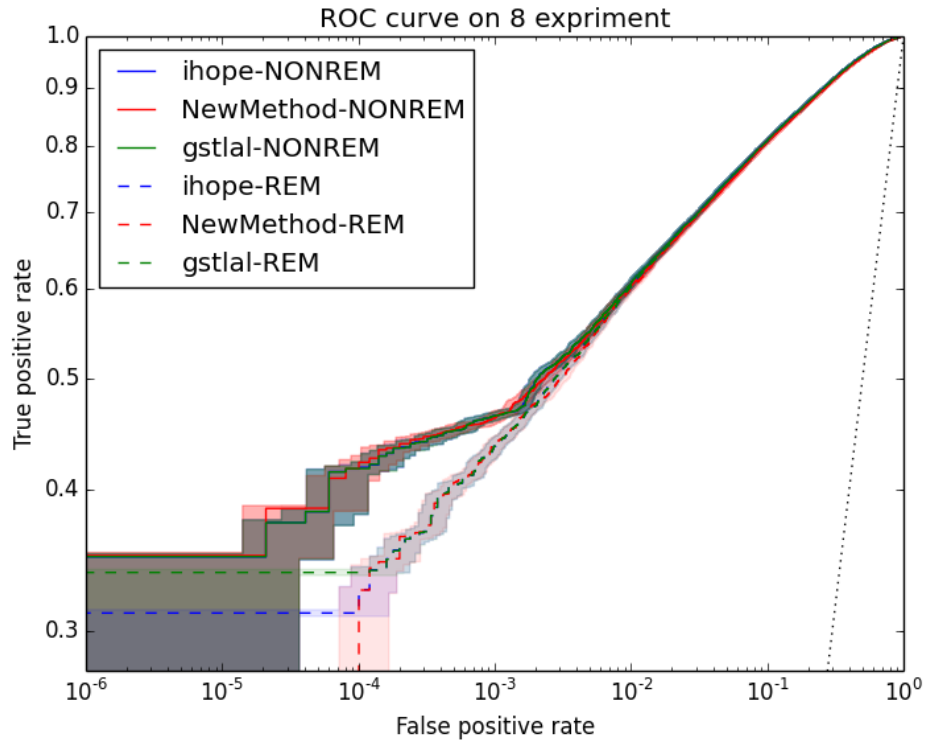


Figure 6.12: ROC plot for experiment 8. The error bars in both horizontal and vertical direction are calculated with a binomial likelihood under a 68.3% credible interval. For each point, the outer error-bar of horizontal and vertical will be picked up to be shown. The solid lines correspond to non-removals while dashed lines removal.

### Medium foreground rate

Experiments 6, 8, 9 and 13 contain medium foreground rates and collectively show 2 types of behaviour. Experiments 9 and 13 show general agreement between algorithms and removal and non-removal approaches. Plots of these ROC curves are shown in Figs. A.26 and A.29. Experiments 6 and 8 show similar deviations to those seen in the  $p$ - $p$  plots in Section 6.3.2. This similarity is not surprising since the vertical axes of the  $p$ - $p$  plots are identical to horizontal axes of the ROC plots with the distinction that they are computed on background-only and background-foreground experiments respectively. Here we focus on Experiment 8 shown in Fig. 6.12 which contained realistic but slightly different background distributions in each detector. As seen in the low-foreground example there is good agreement between algorithms but differences between removal and non-removal approaches. In this case, due to the increased number of foreground events, this difference is more clearly visible and the discrepancies are significant at the multiple- $\sigma$  level. So for medium foreground rates we conclude that as a detection statistic, those methods that apply non-removal obtain higher TPRs at fixed FPR for low values of FAP. We remind the reader that detection claims will be made at low FAP values.

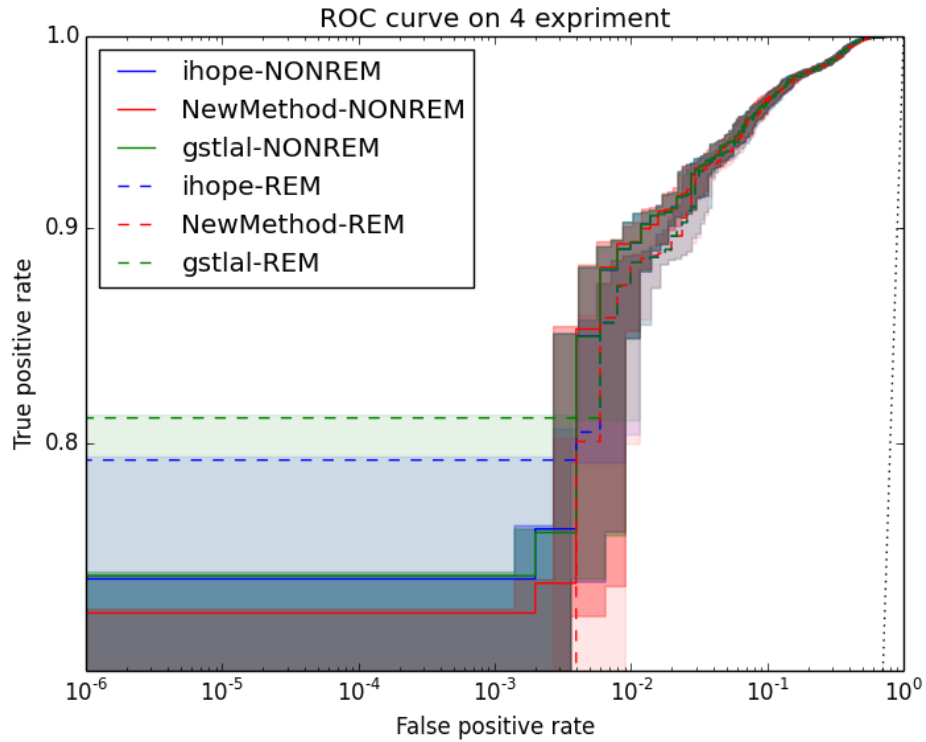


Figure 6.13: ROC plot for experiment 4. The error bars in both horizontal and vertical direction are calculated with a binomial likelihood under a 68.3% credible interval. For each point, the outer error-bar of horizontal and vertical will be picked up to be shown. The solid lines correspond to non-removals while dashed lines removal.

### High foreground rate

The high rate experiments 4, 5 and 11 all show similar behaviour. Here we show Fig. 6.13 as an example where we see general agreement (within our estimated uncertainties) between algorithms and removal and non-removal. The high rates used result in  $>90\%$  of realisations containing a loudest coincident event from the foreground. The 3 experiments were examples of all 3 levels of background complexity respectively and the results indicate that in terms of detection efficiency, all algorithms and approaches perform equally well at any fixed FPR. This high rate scenario is most likely to be relevant in the epochs following the first direct detection. As the CDF distribution of the foreground signals are much flatter than the fast dropping background distribution, so that when the foreground rate is high enough, the SNR corresponding to, say, FPR  $10^{-2}$  will immediate means that in any realisation, it'll be extremely difficult for any background triggers to reach that high SNR

### Detection efficiency

We have taken slices through the ROC plots at selected fixed values of FPR and extracted the TPR and its associated uncertainty. We have done this for all experiments containing a non-zero foreground rate and listed in results in Table 6.2. We have highlighted in bold those methods that achieved the highest detection efficiency for each FPR where results differed by greater than their estimated uncertainty.

Our MDC contained a limited number of experiment realisations, so the rarest event we would expect to generate from all of our background distributions would have  $\text{FAP} \sim 10^{-5}$ . We have therefore set fixed FPRs at the values shown in the table, going no lower than  $4\sigma \equiv 6.3 \times 10^{-5}$  and no higher than  $10^{-2}$ , since all methods tend to agree at these values and higher, and  $4\sigma$  is still not subject to random fluctuation too much. In total we found there are 34 occurrences that in terms of having higher detection efficiency favour non-removal methods, while in 20 cases removal is favoured. Notice that there are 6 cases that non-removal performs as well as removal. We find that with the exception of experiment 13, all experiments in which removal is favoured either have an extreme background distribution, or a high foreground rate, or a combination of both. Experiment 13 has a very simple background and a low foreground rate and in this case removal methods give higher detection efficiencies, however the difference is not large.

#### 6.3.4 Boxplots

In this section we look at decade width slices in true FAP and show the defining features of the corresponding distributions of estimated values. For a given experiment we first take each algorithm and removal method we isolate those results corresponding to the chosen FAP decade and take the ratio of estimated to exact FAP value. The corresponding boxplot then consists of the central box, median point, whiskers and individual outliers. The central box contains the central half of all points, which starts from 25% to 75% of the total sorted values (known as the first and the third quartile), the length of the box is also known as the inter-quartile range (IQR). The box is divided by a vertical line identifying the 50% (median) quartile. When the distribution is relatively concentrated, and the most extreme samples lie within  $1.5 \times$  the IQR, then the whisker ends at the most extreme point. Otherwise the whisker is drawn to be  $1.5 \times$  the IQR, and outliers beyond the whisker are drawn as individual points. We also indicate on these plots the mean value of (non-log) ratio for each distribution.

Since we are more interested in the region of low FAP values, where detection of foreground signals will likely occur, we have decided to slice data into the following ranges:  $(10^{-5} - 10^{-4})$ ,  $(10^{-4} - 10^{-3})$ , and  $(10^{-3} - 10^{-2})$ . For each data bin, we draw boxplots on the ratio between the estimated and exact FAP value and use a logarithmic x-axis scale. In this case

True positive rate (%)		False positive rate				
Experiment	Method	$4\sigma$	$10^{-4}$	$10^{-3}$	$3\sigma$	$10^{-2}$
10	ihope non-rm	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$
	ihope rm	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$
	New Method non-rm	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$46^{+13}_{-12}$
	New Method rm	$15^{+15}_{-5.4}$	$23^{+15}_{-7.6}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$46^{+13}_{-12}$
	gstlal non-rm	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$
2	gstlal rm	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$	$38^{+14}_{-11}$
	ihope non-rm	$39.3^{+3.5}_{-3.3}$	$40.3^{+3.5}_{-3.3}$	$44.3^{+3.5}_{-3.4}$	$53.2^{+3.5}_{-3.5}$	$61.2^{+3.3}_{-3.5}$
	ihope rm	$24.4^{+3.3}_{-2.8}$	$27.4^{+3.4}_{-2.9}$	$41.8^{+3.5}_{-3.4}$	$53.2^{+3.5}_{-3.5}$	$61.2^{+3.3}_{-3.5}$
	New Method non-rm	$39.3^{+3.5}_{-3.3}$	$40.3^{+3.5}_{-3.3}$	$45.3^{+3.5}_{-3.4}$	$53.2^{+3.5}_{-3.5}$	$63.2^{+3.3}_{-3.5}$
	New Method rm	$16.4^{+2.9}_{-2.3}$	$22.3^{+3.2}_{-2.7}$	$41.8^{+3.5}_{-3.4}$	$53.2^{+3.5}_{-3.5}$	$61.7^{+3.3}_{-3.5}$
7	gstlal non-rm	$39.3^{+3.5}_{-3.3}$	$40.3^{+3.5}_{-3.3}$	$44.3^{+3.5}_{-3.4}$	$53.7^{+3.5}_{-3.5}$	$61.2^{+3.3}_{-3.5}$
	gstlal rm	$32.8^{+3.5}_{-3.1}$	$32.8^{+3.5}_{-3.1}$	$41.8^{+3.5}_{-3.4}$	$53.2^{+3.5}_{-3.5}$	$61.2^{+3.3}_{-3.5}$
	ihope non-rm	$58.5^{+3.2}_{-3.4}$	$58.9^{+3.2}_{-3.4}$	$74.6^{+2.7}_{-3.1}$	$79.0^{+2.5}_{-3.0}$	$85.7^{+2.0}_{-2.7}$
	ihope rm	$62.1^{+3.1}_{-3.3}$	$62.1^{+3.1}_{-3.3}$	$74.1^{+2.7}_{-3.1}$	$79.0^{+2.5}_{-3.0}$	$85.7^{+2.0}_{-2.7}$
	New Method non-rm	$57.6^{+3.2}_{-3.4}$	$59.4^{+3.2}_{-3.4}$	$73.7^{+2.7}_{-3.1}$	$80.4^{+2.4}_{-2.9}$	$86.2^{+2.0}_{-2.6}$
13	New Method rm	$62.1^{+3.1}_{-3.3}$	$62.5^{+3.1}_{-3.3}$	$75.6^{+2.7}_{-3.1}$	$80.4^{+2.4}_{-2.9}$	$86.2^{+2.0}_{-2.6}$
	gstlal non-rm	$58.5^{+3.2}_{-3.4}$	$59.4^{+3.2}_{-3.4}$	$74.6^{+2.7}_{-3.1}$	$79.0^{+2.5}_{-3.0}$	$85.7^{+2.0}_{-2.7}$
	gstlal rm	$62.1^{+3.1}_{-3.3}$	$62.1^{+3.1}_{-3.3}$	$74.1^{+2.7}_{-3.1}$	$79.0^{+2.5}_{-3.0}$	$85.7^{+2.0}_{-2.7}$
	ihope non-rm	$59.40^{+0.75}_{-0.76}$	$61.36^{+0.75}_{-0.76}$	$72.20^{+0.68}_{-0.70}$	$76.79^{+0.64}_{-0.67}$	$82.37^{+0.57}_{-0.60}$
	ihope rm	$60.35^{+0.75}_{-0.76}$	$62.22^{+0.74}_{-0.75}$	$72.15^{+0.68}_{-0.70}$	$76.55^{+0.64}_{-0.67}$	$82.33^{+0.57}_{-0.60}$
9	New Method non-rm	$58.75^{+0.76}_{-0.79}$	$60.81^{+0.75}_{-0.76}$	$72.01^{+0.68}_{-0.70}$	<b><math>76.57^{+0.64}_{-0.67}</math></b>	$82.06^{+0.58}_{-0.61}$
	New Method rm	$60.33^{+0.75}_{-0.76}$	<b><math>62.55^{+0.74}_{-0.75}</math></b>	$72.18^{+0.68}_{-0.70}$	$76.45^{+0.64}_{-0.67}$	$81.99^{+0.58}_{-0.61}$
	gstlal non-rm	$59.45^{+0.75}_{-0.76}$	$61.43^{+0.75}_{-0.76}$	$72.20^{+0.68}_{-0.70}$	$76.79^{+0.64}_{-0.67}$	$82.37^{+0.57}_{-0.60}$
	gstlal rm	<b><math>60.38^{+0.75}_{-0.76}</math></b>	$62.22^{+0.74}_{-0.75}$	$72.15^{+0.68}_{-0.70}$	$76.55^{+0.64}_{-0.67}$	$82.30^{+0.57}_{-0.61}$
	ihope non-rm	$59.29^{+0.36}_{-0.36}$	$60.15^{+0.36}_{-0.36}$	$67.94^{+0.34}_{-0.34}$	$72.43^{+0.32}_{-0.33}$	$79.01^{+0.29}_{-0.30}$
8	ihope rm	$57.76^{+0.36}_{-0.36}$	$58.34^{+0.36}_{-0.36}$	$67.26^{+0.34}_{-0.34}$	$71.77^{+0.33}_{-0.33}$	$78.92^{+0.29}_{-0.30}$
	New Method non-rm	$58.93^{+0.36}_{-0.36}$	$59.47^{+0.36}_{-0.36}$	$67.67^{+0.34}_{-0.34}$	$71.95^{+0.33}_{-0.33}$	$78.76^{+0.29}_{-0.30}$
	New Method rm	$58.41^{+0.36}_{-0.36}$	$58.56^{+0.36}_{-0.36}$	$67.07^{+0.34}_{-0.34}$	$71.65^{+0.33}_{-0.33}$	$78.71^{+0.29}_{-0.30}$
	gstlal non-rm	<b><math>59.32^{+0.36}_{-0.36}</math></b>	<b><math>60.22^{+0.35}_{-0.36}</math></b>	<b><math>67.97^{+0.34}_{-0.34}</math></b>	<b><math>72.44^{+0.32}_{-0.33}</math></b>	$79.00^{+0.29}_{-0.30}$
	gstlal rm	$57.76^{+0.36}_{-0.36}$	$58.34^{+0.36}_{-0.36}$	$67.27^{+0.34}_{-0.34}$	$71.76^{+0.33}_{-0.33}$	$78.92^{+0.29}_{-0.30}$
6	ihope non-rm	$41.42^{+0.22}_{-0.22}$	$41.66^{+0.22}_{-0.22}$	$46.41^{+0.22}_{-0.22}$	<b><math>51.44^{+0.22}_{-0.22}</math></b>	$60.56^{+0.22}_{-0.22}$
	ihope rm	$30.16^{+0.21}_{-0.20}$	$30.16^{+0.21}_{-0.20}$	$43.63^{+0.22}_{-0.22}$	$49.92^{+0.22}_{-0.22}$	$60.35^{+0.22}_{-0.22}$
	New Method non-rm	$40.88^{+0.22}_{-0.22}$	<b><math>41.69^{+0.22}_{-0.22}</math></b>	$46.21^{+0.22}_{-0.22}$	$50.92^{+0.22}_{-0.22}$	$60.33^{+0.22}_{-0.22}$
	New Method rm	$14.52^{+0.16}_{-0.16}$	$24.45^{+0.19}_{-0.19}$	$43.75^{+0.22}_{-0.22}$	$49.09^{+0.22}_{-0.22}$	$59.97^{+0.22}_{-0.22}$
	gstlal non-rm	$41.40^{+0.22}_{-0.22}$	<b><math>41.69^{+0.22}_{-0.22}</math></b>	<b><math>46.42^{+0.22}_{-0.22}</math></b>	$51.43^{+0.22}_{-0.22}$	<b><math>60.57^{+0.22}_{-0.22}</math></b>
5	gstlal rm	$33.79^{+0.21}_{-0.21}$	$33.79^{+0.21}_{-0.21}$	$43.65^{+0.22}_{-0.22}$	$49.93^{+0.22}_{-0.22}$	$60.34^{+0.22}_{-0.22}$
	ihope non-rm	$65.99^{+0.18}_{-0.19}$	$66.47^{+0.18}_{-0.19}$	$71.91^{+0.17}_{-0.18}$	<b><math>80.25^{+0.15}_{-0.16}</math></b>	$87.18^{+0.13}_{-0.13}$
	ihope rm	$55.03^{+0.19}_{-0.19}$	$55.03^{+0.19}_{-0.19}$	$69.99^{+0.18}_{-0.18}$	$79.36^{+0.16}_{-0.16}$	$87.26^{+0.13}_{-0.13}$
	New Method non-rm	<b><math>66.95^{+0.18}_{-0.18}</math></b>	<b><math>67.15^{+0.18}_{-0.18}</math></b>	$72.87^{+0.17}_{-0.17}$	<b><math>79.75^{+0.16}_{-0.16}</math></b>	$86.97^{+0.13}_{-0.13}$
	New Method rm	$26.35^{+0.17}_{-0.17}$	$26.79^{+0.17}_{-0.17}$	$69.92^{+0.18}_{-0.18}$	$78.41^{+0.16}_{-0.16}$	$86.84^{+0.13}_{-0.13}$
4	gstlal non-rm	$66.04^{+0.18}_{-0.19}$	$66.48^{+0.18}_{-0.18}$	$71.89^{+0.17}_{-0.18}$	$80.24^{+0.15}_{-0.16}$	$87.23^{+0.13}_{-0.13}$
	gstlal rm	$60.29^{+0.19}_{-0.19}$	$60.29^{+0.19}_{-0.19}$	$70.00^{+0.18}_{-0.18}$	$79.37^{+0.16}_{-0.16}$	$87.27^{+0.13}_{-0.13}$
	ihope non-rm	$94.28^{+0.07}_{-0.07}$	$94.28^{+0.07}_{-0.07}$	$94.28^{+0.07}_{-0.07}$	$95.09^{+0.07}_{-0.07}$	$98.08^{+0.04}_{-0.05}$
	ihope rm	$91.04^{+0.09}_{-0.09}$	$91.04^{+0.09}_{-0.09}$	$91.04^{+0.09}_{-0.09}$	$93.06^{+0.08}_{-0.08}$	$98.10^{+0.04}_{-0.04}$
	New Method non-rm	$94.28^{+0.07}_{-0.08}$	$94.28^{+0.07}_{-0.08}$	$94.28^{+0.07}_{-0.08}$	<b><math>95.16^{+0.07}_{-0.07}</math></b>	$97.16^{+0.05}_{-0.05}$
11	New Method rm	$39.20^{+0.15}_{-0.15}$	$39.20^{+0.15}_{-0.15}$	$39.20^{+0.15}_{-0.15}$	$93.43^{+0.08}_{-0.08}$	$97.24^{+0.05}_{-0.05}$
	gstlal non-rm	<b><math>94.30^{+0.07}_{-0.07}</math></b>	<b><math>94.30^{+0.07}_{-0.07}</math></b>	<b><math>94.30^{+0.07}_{-0.07}</math></b>	$95.09^{+0.07}_{-0.07}$	$98.08^{+0.04}_{-0.04}$
	gstlal rm	$93.07^{+0.08}_{-0.08}$	$93.07^{+0.08}_{-0.08}$	$93.07^{+0.08}_{-0.08}$	$93.07^{+0.08}_{-0.08}$	$98.10^{+0.04}_{-0.04}$
	ihope non-rm	$74.25^{+0.14}_{-0.14}$	$74.25^{+0.14}_{-0.14}$	$74.25^{+0.14}_{-0.14}$	$76.31^{+0.13}_{-0.14}$	$89.28^{+0.10}_{-0.10}$
	ihope rm	$78.49^{+0.13}_{-0.13}$	$78.49^{+0.13}_{-0.13}$	$78.49^{+0.13}_{-0.13}$	$79.25^{+0.13}_{-0.13}$	$88.32^{+0.10}_{-0.10}$
1	New Method non-rm	$72.87^{+0.14}_{-0.14}$	$72.87^{+0.14}_{-0.14}$	$72.87^{+0.14}_{-0.14}$	$74.07^{+0.14}_{-0.14}$	<b><math>89.34^{+0.10}_{-0.10}</math></b>
	New Method rm	$15.58^{+0.12}_{-0.11}$	$15.58^{+0.12}_{-0.11}$	$15.58^{+0.12}_{-0.11}$	$43.45^{+0.16}_{-0.16}$	$88.32^{+0.10}_{-0.10}$
	gstlal non-rm	$74.38^{+0.14}_{-0.14}$	$74.38^{+0.14}_{-0.14}$	$74.38^{+0.14}_{-0.14}$	$76.16^{+0.13}_{-0.14}$	$89.31^{+0.10}_{-0.10}$
	gstlal rm	<b><math>81.10^{+0.12}_{-0.12}</math></b>	<b><math>81.10^{+0.12}_{-0.12}</math></b>	<b><math>81.10^{+0.12}_{-0.12}</math></b>	<b><math>81.10^{+0.12}_{-0.12}</math></b>	$88.32^{+0.10}_{-0.10}$
	ihope non-rm	$93.94^{+0.08}_{-0.08}$	$93.94^{+0.08}_{-0.08}$	$93.94^{+0.08}_{-0.08}$	$94.46^{+0.07}_{-0.07}$	$95.32^{+0.07}_{-0.07}$
11	ihope rm	<b><math>94.64^{+0.07}_{-0.07}</math></b>	<b><math>94.64^{+0.07}_{-0.07}</math></b>	<b><math>94.64^{+0.07}_{-0.07}</math></b>	$94.71^{+0.07}_{-0.07}$	$95.25^{+0.07}_{-0.07}$
	New Method non-rm	$93.66^{+0.08}_{-0.08}$	$93.66^{+0.08}_{-0.08}$	$93.66^{+0.08}_{-0.08}$	$93.78^{+0.08}_{-0.08}$	<b><math>96.00^{+0.06}_{-0.06}</math></b>
	New Method rm	$94.13^{+0.07}_{-0.08}$	$94.13^{+0.07}_{-0.08}$	$94.13^{+0.07}_{-0.08}$	$94.49^{+0.07}_{-0.07}$	$95.20^{+0.07}_{-0.07}$
	gstlal non-rm	$93.96^{+0.08}_{-0.08}$	$93.96^{+0.08}_{-0.08}$	$93.96^{+0.08}_{-0.08}$	$94.48^{+0.07}_{-0.07}$	$95.34^{+0.07}_{-0.07}$
	gstlal rm	<b><math>94.64^{+0.07}_{-0.07}</math></b>	<b><math>94.64^{+0.07}_{-0.07}</math></b>	<b><math>94.64^{+0.07}_{-0.07}</math></b>	<b><math>94.72^{+0.07}_{-0.07}</math></b>	$95.25^{+0.07}_{-0.07}$

Table 6.2: Detection efficiency (in percent) at given FPR, together with the uncertainty considering a binomial likelihood. For each FPR, there are 6 values for detection efficiency, we identify the biggest of them, and find the corresponding category of either removal or non-removal. If this value differ the highest detection efficiency in the other category by at least one error bar, we set such value bold font.

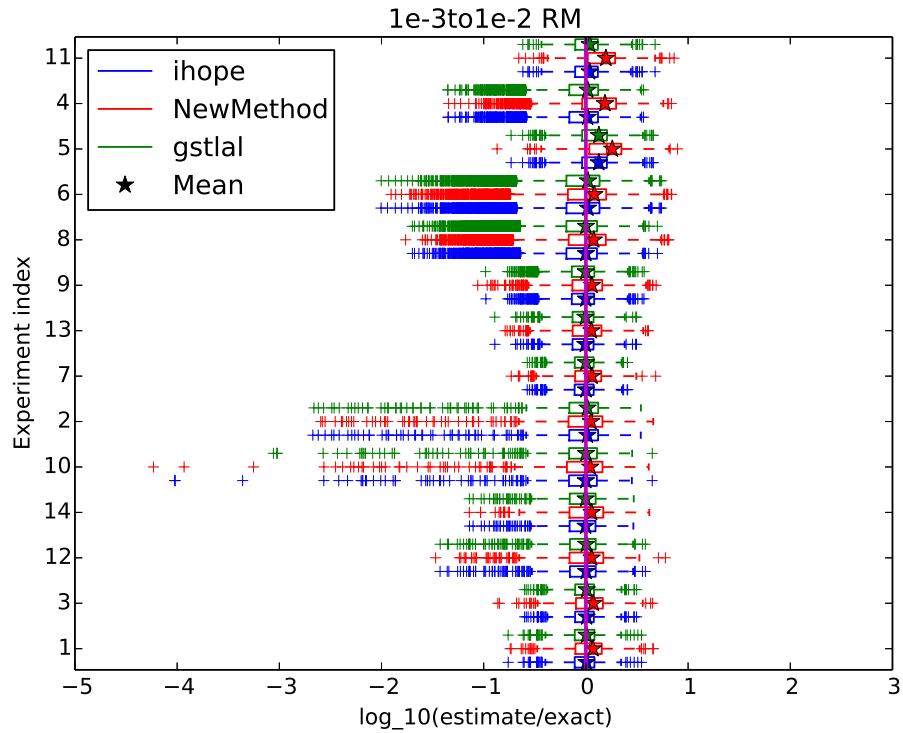
a factor of 10 overestimation appears as far to the right hand side of the plot as a factor of 10 underestimation does to the left. The vertical purple line corresponds to the log of ratio is zero, which means the estimation and exact FAP is identical; left hand side means the estimated FAP value is smaller then the actual value, which translates to an overestimation of FAP. In all plots the vertical axis gives the experiment index ranging from the lowest foreground rates to the highest and for each index there are 3 coloured boxes associated with each algorithm. Figures are divided into 2 plots, one for the removal methods and the other for non-removal.

So if we have a large number of detections, by applying the removal method we know that on average the estimation of FAP is unbiased. However, we also know that the individual estimations are much more likely to be overestimated rather than underestimated.

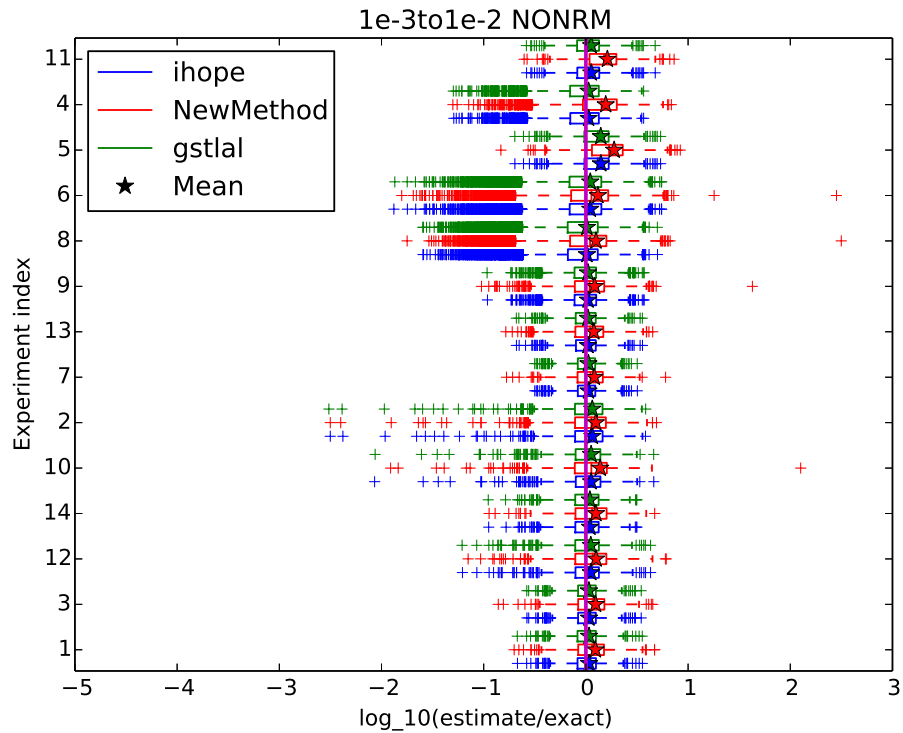
### **FAP range $10^{-3}$ – $10^{-2}$**

In Fig. 6.14 we see relatively tight and symmetric distributions for all algorithms when considering the IQR with strong agreement specifically between the gstlal and ihope. We remind the reader that the New Method was not optimised at high FAP values and hence shows very slightly broader distributions. We note that the extrema of the FAP ratios in most experiments range symmetrically from  $\sim \pm 1$  order of magnitude. However, for some experiments, most notably 4, 6, 8, 2, 10, and 12 there are large deviations in the extrema to particularly sizeable underestimates of FAP. This tendency would be classed as a non-conservative effect whereby events could be estimated as more rare than they truly are. This effect is partially reduced in the case of non-removal approaches as is most evident for experiments 2 and 10. All of these experiments for which there are sizeable underestimates have either realistic or extreme background distributions. However, such backgrounds do not give underestimates exclusively since experiment 11 shows no such behaviour. We notice that some experiments show similarity in the extrema distribution, like experiment 2 and 10, experiment 6 and 8. They don't share the same background feature, like experiment 2 have identical background distributions for the two detectors, which contains a platform inside. While for experiment 10, there's no platform feature, but the two detectors are quite asymmetric. What they share for similarity is only the foreground event rates, and the fact that some background triggers, although they do not form actual coincidence, could generate very loud artificial coincidences and thus fool the estimation.

The points identified with star symbols in Fig. 6.14 show the means of the distribution of ratios (this is not the mean of the log of the ratio). It can be seen that in general, the distribution means for removal methods provide values slightly more consistent with the expected vertical line. As we will show in the subsequent sections as we investigate lower FAP decades this effect becomes amplified. For this ( $10^{-3}$ – $10^{-2}$ ) region we note that for the same reasons



(a) Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-3}$  and  $10^{-2}$  based on removal



(b) Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-3}$  and  $10^{-2}$  based on non-removal

Figure 6.14: Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-3}$  and  $10^{-2}$ . Left means optimistic while right is conservative. The upper 3 are with high rate and the bottom 4 are with zero rate.

as discussed earlier, the means computed from the New Method tends to overestimate the expected value.

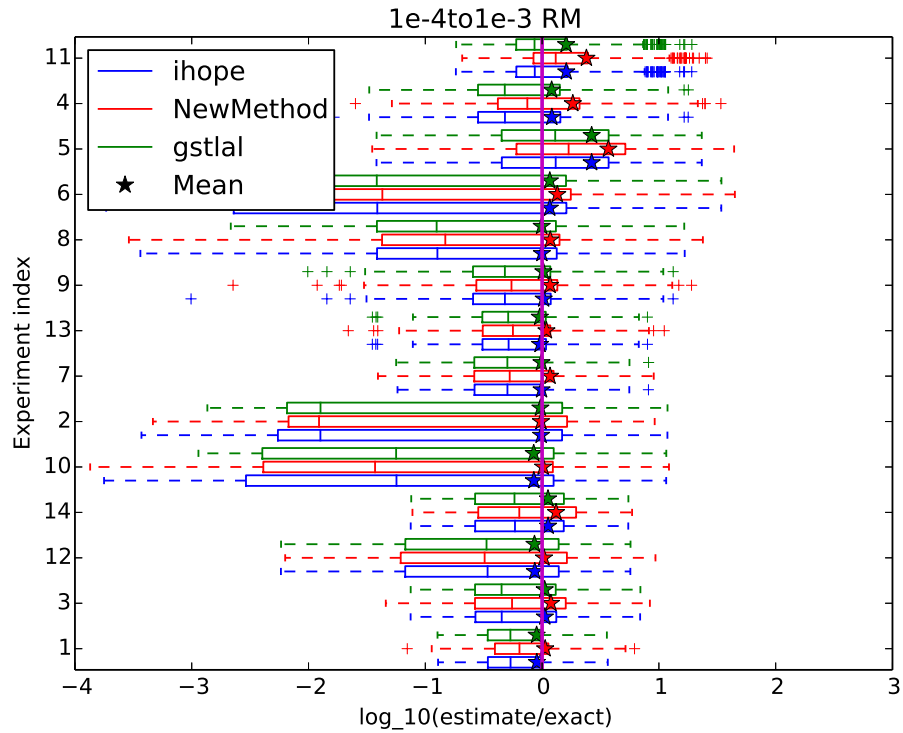
### **FAP range $10^{-4}$ – $10^{-3}$**

As we move to lower FAP ranges, shown in Fig. 6.15, we start to see the effects of having lower numbers of results. By definition we would expect to see a factor of  $\approx 10$  fewer results in the decade ( $10^{-4}$ – $10^{-3}$ ) as compared to the previous section. This implies larger statistical fluctuations due the reduced number of samples but also leads to intrinsically broader distributions. The reason for the latter feature is that the estimation methods themselves are equally constrained by the infrequency of loud, low-FAP events. As seen in previous figures of merit, results differ only slightly between algorithms with the largest differences coming from the issue of removal or non-removal of zero-lag triggers.

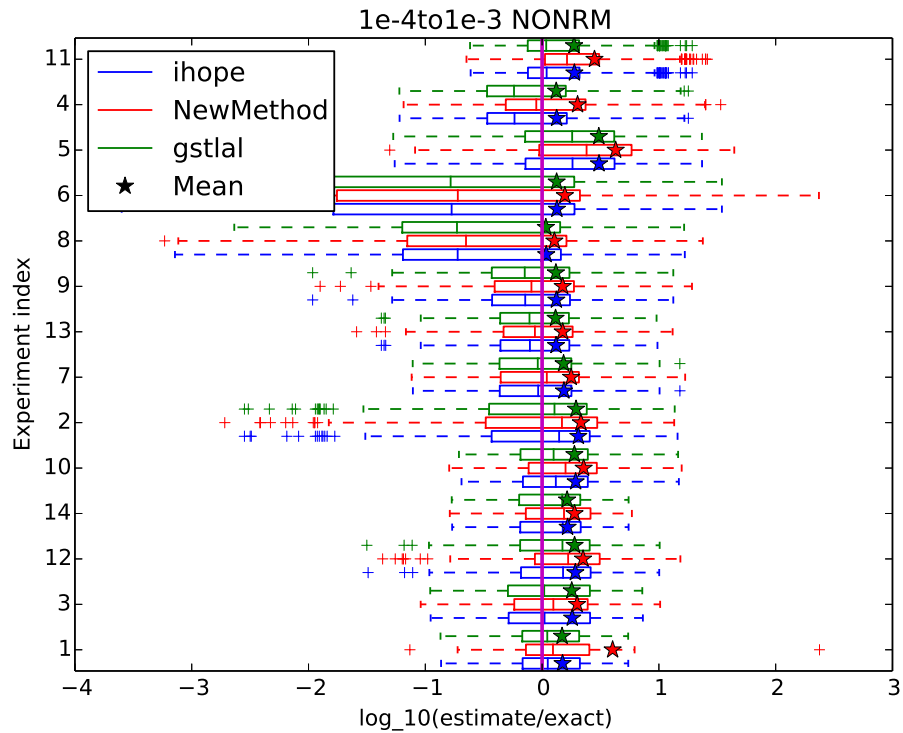
Overall, we see ranges in the extrema spanning  $\pm 1$  order of magnitude for both removal and non-removal methods. However, for experiments 10, 2, 6, and 8 the lower extrema extend to  $\sim 4$  order of magnitude below the expected ratio for the removal method. This behaviour is moderately reduced for the non-removal method where we note that for experiment 10 the extrema are entirely reduced to be consistent with the majority of other experiments. In general it is clear that the IQRs for the removal method are far broader in logarithmic space than their non-removal equivalents. This increase in width is always to lower values of the ratio meaning underestimates of the FAP similarly exemplified by the location of the median values. For removal methods low foreground rates yield medians skewed to lower values by factors of  $\sim 2$ – $200$ . For the 3 high foreground rate experiments the IQRs and corresponding medians appear consistent with the expected values. For the non-removal methods the IQRs and medians are relatively symmetric about the expected value and the IQRs are in all cases narrower than for the removal methods.

In this FAP range it becomes clearer that there is a definite difference between the removal and non-removal methods with regards to the distribution means. Removal methods consistently return mean estimates that are well within factors of 2 for all low and medium foreground rates. For high foreground rates they consistently overestimate the means by up to a factor of  $\sim 3$ . For the non-removal cases there is a clear overestimate of the ratio (implying a conservative overestimate of the FAP) for all experiments irrespective of foreground rate or background complexity. This overestimate is in general a factor of  $\sim 2$ . It is interesting to note that the estimates from both approaches for the 3 high foreground rate experiments are very similar in their distributions and means.





(a) Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-4}$  and  $10^{-3}$  based on removal



(b) Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-4}$  and  $10^{-3}$  based on non-removal

Figure 6.15: Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-4}$  and  $10^{-3}$ . Left means optimistic while right is conservative. The upper 3 are with high rate and the bottom 4 are with zero rate.

**FAP range**  $10^{-5}$ – $10^{-4}$ 

In this FAP range the uncertainties and variation in the results are strongly influenced by the low number of events present at such low FAP values. Nonetheless, in Fig. 6.16 we see interesting similarities between algorithms and striking differences between removal and non-removal approaches. Firstly, in all cases the variation in extrema is comparable, in this case spanning  $\sim \pm 3$  orders of magnitude. The IQRs are broadly scattered and in many cases do not intersect with the expected values. This is not indicative of poor estimation but indicative of highly non-symmetric distributions.

For low foreground rates there is a marked difference between results from removal and non-removal methods. For removal all distributions are skewed to low values which is also a characteristic for the medium foreground rate experiments. For example, in experiment 12 there are no estimates in any of the realisations in this range that overestimate the FAP. Removal methods in general in this range of very low FAP for low and medium foreground rates provide IQRs of width  $\sim 1$  order of magnitude with medians shifted by between  $\sim 1$ – $2$  orders of magnitude below the expected values. For non-removal methods, all low foreground experiments (with the exception of experiment 2) provide conservative overestimates of the FAP ratio with IQRs and extrema ranges spanning  $< 1$  and  $\sim 1$  order of magnitude respectively. In contrast to experiment 12 with the removal method, we see that for experiment 10 there are no non-removal estimates in any realisation that underestimate the FAP. With non-removal there is then a marked change as we move to medium level foreground rates and the distributions become relatively symmetric in log-space with all medians lower than, but mostly within a factor of 2 of, the expected value. Experiments 6 and 8 both have medium level foreground rates and give rise to results that are very similar between removal and non-removal results and that exhibit highly skewed distributions to low values with long distribution tails extending to high values. This trend of similarity is then continued for high foreground rates where there is little distinction between either algorithm of removal methods. In these cases however, the distributions appear relatively symmetric in log-space with reasonably well constrained IQRs.

If we focus on the estimates of the means of the distributions then we see similar behaviour but with more variation than in the previous FAP ranges. Starting with non-removal methods there is a consistent conservative bias in the mean of the ratio of estimated to true FAP. For low and high foreground rates this bias is  $\sim 1$  order of magnitude which reduces to a factor of  $\sim 3$  overestimate for medium level foregrounds. For the removal methods, low foreground rates return distribution means that are scattered symmetrically about the expected value with a variation of  $\sim 1$  order of magnitude. For all medium level foregrounds including experiments with low, medium and high background complexity, the mean estimates are very tightly localised around the expected values with variations of 10's of percent. For high

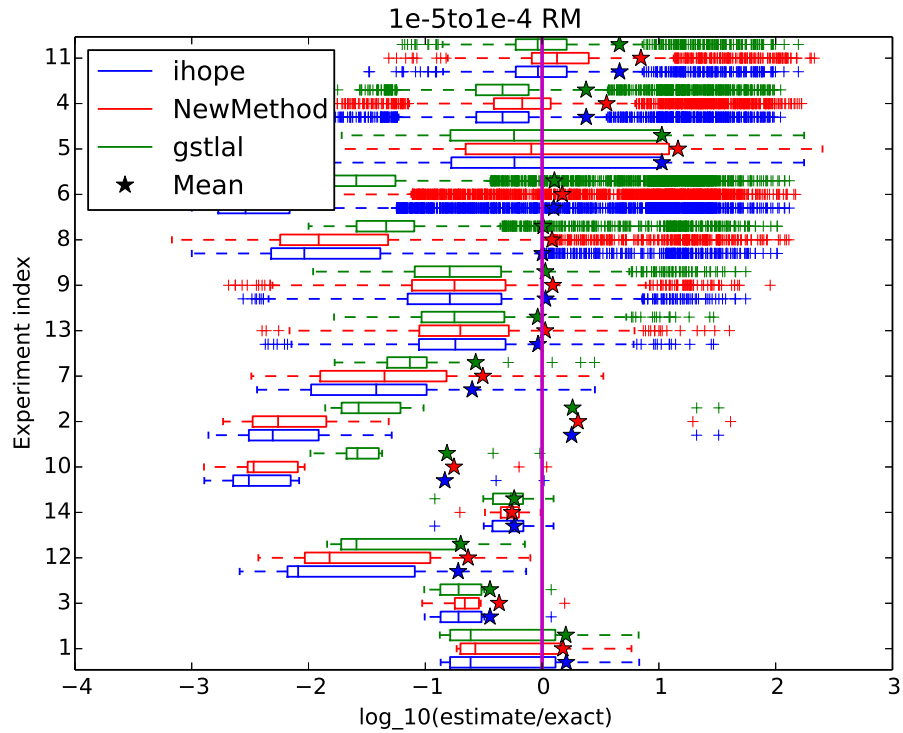
foreground rates the means obtained from both removal and non-removal methods are all consistently overestimates of the expected value by up to  $\sim 1$  order of magnitude.

By looking at the bulk distribution in the boxplots, it seems that removal will generally overestimate the FAP, while although non-removal is no better in precision for individual experiments and even worse in accuracy, systematically the mean value is unbiased over all 14 experiments. However, notice that if we only look at the log of mean, then non-removal almost always underestimate the FAP, while removal generally being consistent with the exact FAP. This indirectly proves that removal method is an unbiased estimator in linear space. For significant event, the exact FAP is very close to zero, any difference due to overestimation will be very small in linear space (although not necessarily small in log space), while underestimation could bias the value relatively hugely. In order to compensate the relatively large bias caused by underestimation to make the removal method still unbiased, the majority of the estimation will have to overestimate.

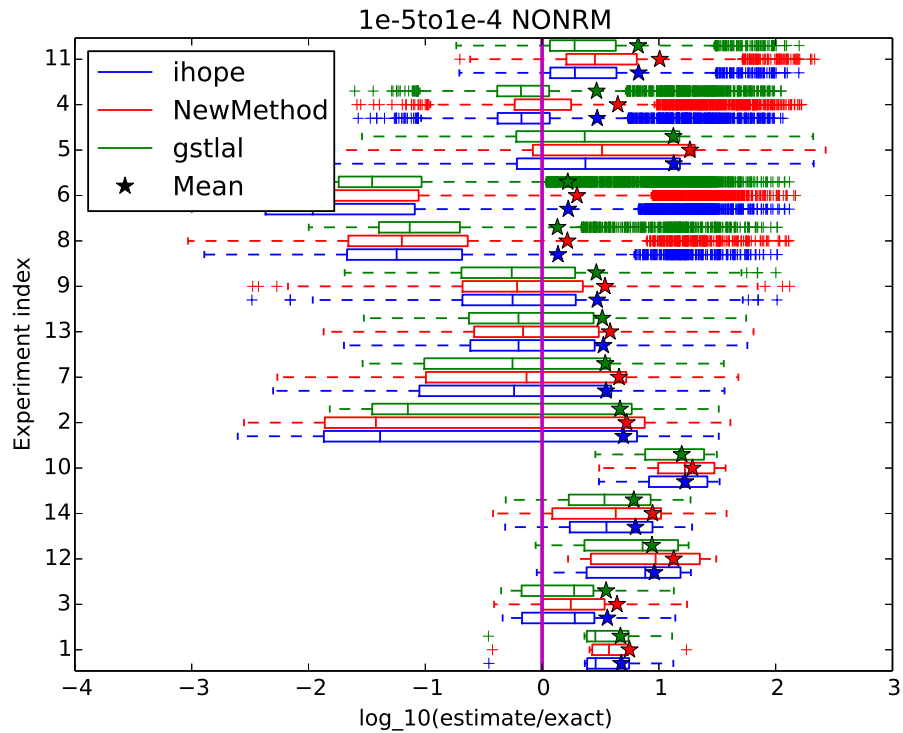
### 6.3.5 Uncertainty in estimation

From the results presented in the previous sections we can conclude that the relative uncertainty in the estimation of FAP increases as FAP decreases. As shown in Figs. 6.2, 6.3, 6.5, and 6.6, with the exception of results from the New Method, which are designed to be accurate only at low FAP values, both other estimation methods show less spread as the FAP value goes higher. Specific features in the background distributions would vary the actual spread, but the tendency is consistent. When the exact FAP is as small as  $10^{-4}$ , the relative uncertainty can exceed 100%, in some cases, as seen in Fig. 6.16, estimates can under or overestimate FAP by many orders of magnitude.

Any claims of GW detection will necessarily be made at low values of FAP and most likely at medium level foreground rate. Using Fig. 6.16 and focussing on the medium foreground rate experiments 10, 2, 7, 13, and 9 it is clear from both removal and non-removal methods that a single loudest event measurement of FAP will be drawn from a very broad distribution. For non-removal methods, in all but experiment 10 for medium foregrounds, the IQR is roughly symmetric and of a width resulting in a 50% probability that a single measurement is within the region described by the true value  $\pm 1$  order of magnitude. For the equivalent removal cases the same IQR (accounting for 50% of the probability) is  $\sim 1/2$  of the width in log-space, and hence more precise. The extrema between approaches are comparable but the bulk of the distribution is more concentrated in the removal case.



(a) Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-5}$  and  $10^{-4}$  based on removal



(b) Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-5}$  and  $10^{-4}$  based on non-removal

Figure 6.16: Boxplot for log of ratio between estimation and exact, drawn on data with FAP between  $10^{-5}$  and  $10^{-4}$ . Left means optimistic while right is conservative. The upper 3 are with high rate and the bottom 4 are with zero rate.

## 6.4 Discussion

I have designed and generated an MDC and compared the results obtained via 3 different algorithms each operating in 2 different modes: estimating the background distributions using either removal or non-removal of zero-lag coincidences. These background distribution estimates were then used to associate a FAP to the loudest coincident event in each realisation of each experiment in the MDC. Our methods for comparison have involved direct comparison of estimates with the true value of FAP, self-consistency checks via the use of  $p$ - $p$  plots for those experiments not containing any foreground signals, an ROC analysis to identify detection efficiency at fixed FPR, and finally a box-plot comparison of the result distributions from each experiment. Based on these comparison analyses we have arrived at the following key conclusions:

- a. Among all three methods of FAP estimation, there is very good agreement given a mode of operation e.g., removal or non-removal of zero-lag triggers. Major discrepancies occur only when comparing results between different operation modes. There is however, good agreement between removal and non-removal approaches for values of FAP.
- b. For low and medium foreground rates and low FAP values, removal approaches tend to underestimate the FAP but there is no evidence to suggest that the mean values of their results are biased.
- c. For low foreground rates and low FAP values, non-removal approaches consistently overestimate the FAP with strong evidence to suggest that the mean of their results is also biased to higher, more conservative values.
- d. For medium level foreground rates the non-removal method has a small bias towards underestimating the FAP in terms of the median of the distributions which becomes apparent at very low FAP values. The mean FAP is always bias to conservative values.
- e. Both removal and non-removal methods perform very similarly for high foreground rates. Median values are spread around the true value and the mean itself is consistently overestimated especially for low FAP values.
- f. According to our design definition of background complexity we do see that experiments that return more extremely low underestimates of FAP come from either complex or realistic distributions and never from simple background models.
- g. In general, either tail or asymmetry between detectors would induce problematic estimation for removal method, as shown in ROC plots,  $p$ - $p$  plots and boxplots, especially when the foreground rate is not high. The estimation for loud events will generally underestimate the FAP, causing a non-conservative estimation.

- h. We have identified the presence of a systematic overestimation of FAP due to the existence of uncoupled loud triggers, where an astrophysical signal has appeared above threshold in one detector but not in the other. Similar effects can be obtained when there is a coincidence in which one astrophysical trigger is very loud while the other barely passes the threshold to be recorded. This situation is physical and could happen in actual data.
- i. The definition of FAP is more self-consistent for non-removal methods. The removal method would claim a fraction of  $10^{-4}$  to have FAP  $10^{-5}$ , so the value of FAP doesn't follow its literal definition. This deviation becomes significant for values of  $\text{FAP} \leq 10^{-3}$ .
- j. In general, FAP estimates computed using non-removal were identified as being better at distinguishing foreground from background at fixed FPR. This was most notable in experiment 8 which contained medium foreground and realistic background. Removal methods were better only when either the background complexity was high or the foreground rate was high.
- k. Removal methods have the merit of appearing to be unbiased estimators in terms of the mean of the FAP estimates. However, the distributions of FAP events are highly asymmetric, especially for low FAP values. Single realisations from removal methods are consequently highly likely to have underestimated values of FAP. Non-removal methods are bias in terms of their mean but, for low FAP events, are highly likely to overestimate the FAP and hence be conservative.
- l. The relative uncertainty in the estimation is larger when the FAP is smaller. The relative uncertainty reaches 100% when the FAP is about  $10^{-4}$  for this MDC. This value depends on the total number of realisations in the experiment and the number of single detector triggers.

At the time of writing this chapter we eagerly await the advanced detector era and have therefore yet to detect any GW signals from an astrophysical foreground. While we are aiming to make concrete detections via the upgraded advanced facilities, we should bear in mind that the first detection would in general be relatively quiet. In this case, we recommend a sacrifice in accuracy of FAP estimation in favour of conservatism. We therefore promote the use of non-removal methods of any of our 3 algorithms thereby anticipating that our results be conservatively overestimated rather than underestimated.

Notice that in the future we expect to make a lot more detections, the interest would shift from estimating FAP from one single loudest event into the estimation astronomical event rate, so we anticipate to set up another Mock Data Challenge for the rate estimation. Such new MDC would be in principle similar to the FAP MDC in terms of the set up, while more physical information would need to be taken into account. One should also notice that as the

detector is getting more sensitive, the recommended approach of FAP estimation can also possibly evolve.<sup>6</sup>

---

<sup>6</sup>This work was done under the collaboration of multiple researchers, with the set up of the Mock Data and the analysis done by myself.

## Chapter 7

# Conclusion and Future work

Through some examples, the importance of applying state-of-the-art statistical methods to gravitational wave data, is demonstrated in this thesis. Bayesian Inference is a very powerful tool that can perform parameter estimation even with a relatively low SNR, and it has the ability of computing the Bayes Factor, which has the potential to quantify the degree of support from the data for a given model, and thus to better distinguish between models by performing Bayesian Model Selection.

One example that demonstrate the power of Bayesian Model Selection is illustrated on the timing data of pulsar 1E 2259+586. These timing data, and the features contained therein, can be fitted by two competing models: either the successive anti-glitches model or the anti-/normal glitch pair model. By applying a Nested Sampling algorithm to both models and computing the Bayes Factor of each model, we established that the successive anti-glitches model clearly has better support from the data, with a Bayes Factor around 45. As new physical models are proposed to explain the intrinsic mechanism of such timing variations, in the future one can perform Bayesian Parameter Estimation on the physical parameters instead of on timing parameters, which could help us to understand the pulsar physics better. This would have profound effect on GW, as neutron stars are expected to be important sources for CBC and continuous waves as well as detecting channel for PTA.

One drawback of Bayesian Inference methods is their general slow speed, especially for multi-modal posterior. In this thesis a novel method, *mixed MCMC*, is developed to accelerate the sampling procedure in order to perform efficient parameter estimation on multi-modal posterior. By enabling the communication between multiple modes, the sampler can efficiently propose candidates among all modes. This method has the potential to accelerate the parameter estimation time for the CBC data analysis, and we hope to explore the implementation of such algorithms in LAL as future work.

The sampling method of mixed MCMC can be applied to more general problems than just Bayesian Inference. An example is the global optimisation of future generation Gravitational



Wave Detectors. One can convert such an optimisation problem into one of finding the detector network configuration with a high Figure of Merit. Such a problem is intrinsically multi-modal and high in dimension, and mixed MCMC becomes an excellent candidate to solve it efficiently. To do this we introduced a *flexibility index* to characterise the number of high Figure of Merit networks that contains one component at a given site. Application of our mixed MCMC approach allowed us to simultaneously optimise the parameters of 3- and 5-detector networks. Our results demonstrate that Australia hosts the best site in terms of having highest flexibility index, either for a 3-detector network or for a 5-detector network. In future work we note that more factors can be introduced into the Figure of Merit, so that one can make the model more realistic or better tuned to specific scientific goals, and the exclusion map used in our analysis can be introduced by considering more realistic constraints – and even allowing our optimisation procedure to be adjusted dynamically.

The slow speed of Bayesian Inference can become the bottleneck of Gravitational Wave data analysis. The detection pipelines like *gstlal* and *ihope* use information on the probability that one trigger event corresponds to the noise to characterise its significance. The process of background estimation can only be approximated, and debate exists for whether or not to remove the foreground for the purpose of estimating background. A Mock Data Challenge was thus set up to test the estimation of significance from both modes. Generally the non-removal method was found to be more conservative, while the removal method is an unbiased estimator for the mean value. However, these methods can only estimate the significance of the loudest events, but not for multiple events. In the future, a similar Mock Data Challenge is planned in order to test the pipeline's ability to determine the astronomical rate for the CBC events, and to assess the significance of more than one event.

# Appendix A

## Additional MDC results

### A.1 Cumulative Distribution Function of SNR

In this section, we show the reverse CDF distribution of the triggers' SNR. For each experiment, two individual detectors could have different background distributions, but the astronomical foreground distribution is the same. [rephrase that sentence, unclear] In the following pictures, two detectors' background SNR distribution is labelled as red and green colour, while the foreground distribution is labeled blue, and their combination is the black line.

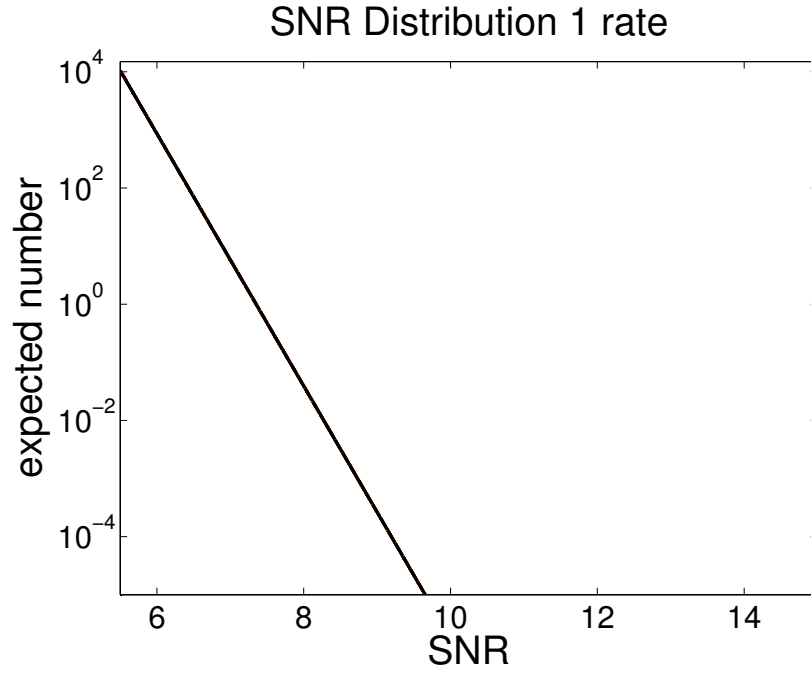


Figure A.1: Reverse CDF distribution of the triggers' SNR for experiment 1. The red and green curves represent the two individual detectors, while the blue curve represents the astronomical signals. The black lines represent the combined distribution of both background and foreground triggers.

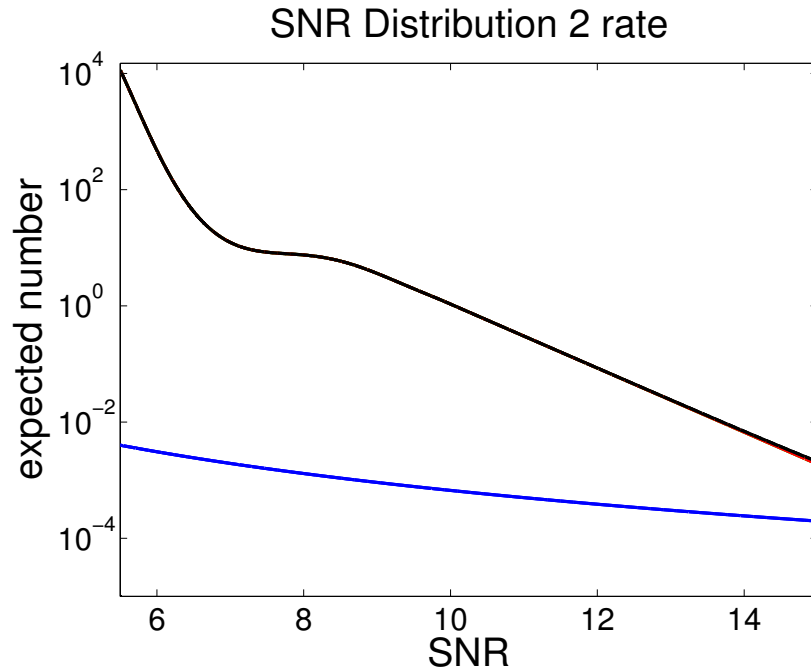


Figure A.2: Reverse CDF of the triggers' SNR for experiment 2: colours assigned as in Fig. A.1.

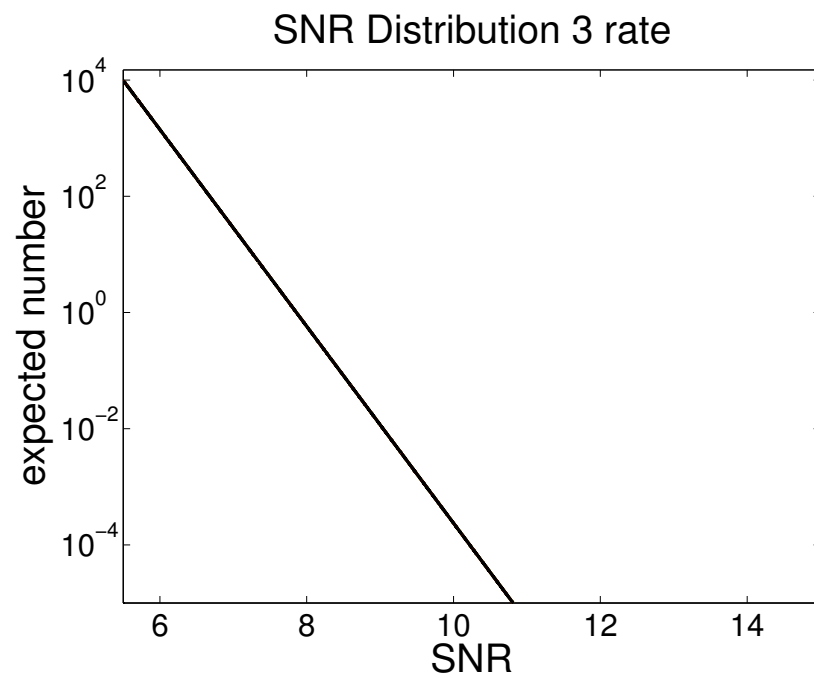


Figure A.3: Reverse CDF of the triggers' SNR for experiment 3: colours assigned as in Fig. A.1.

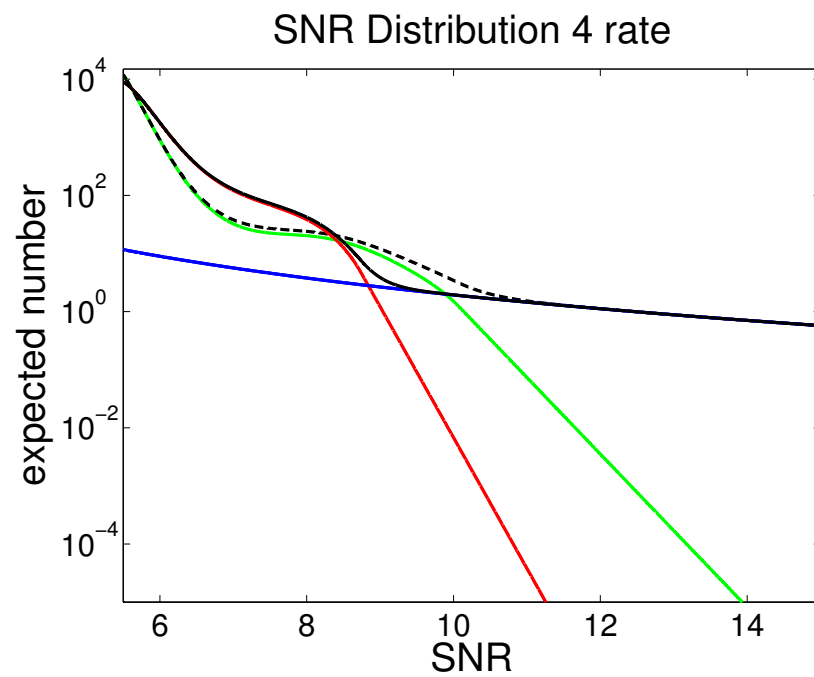


Figure A.4: Reverse CDF of the triggers' SNR for experiment 4: colours assigned as in Fig. A.1.

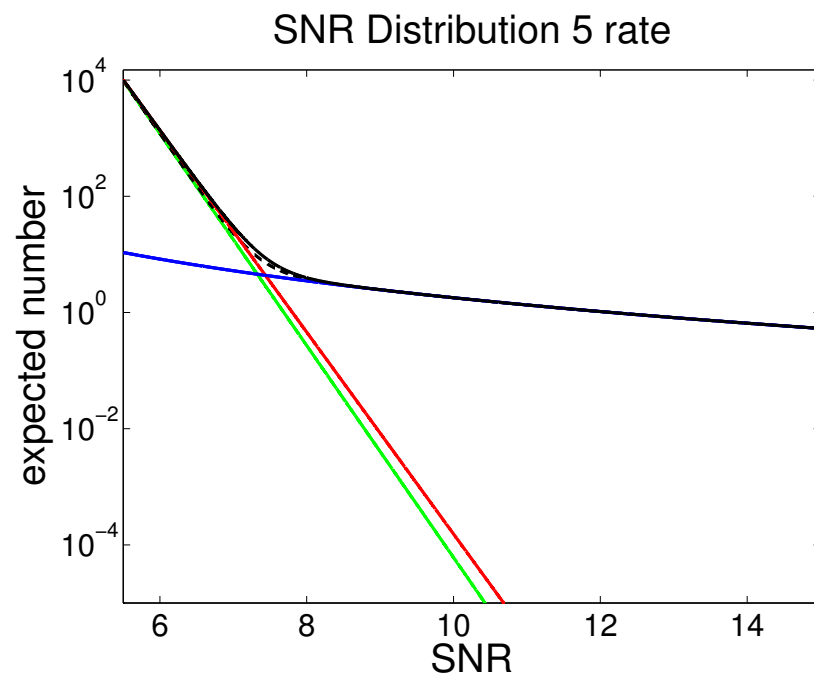


Figure A.5: Reverse CDF of the triggers' SNR for experiment 5: colours assigned as in Fig. A.1.

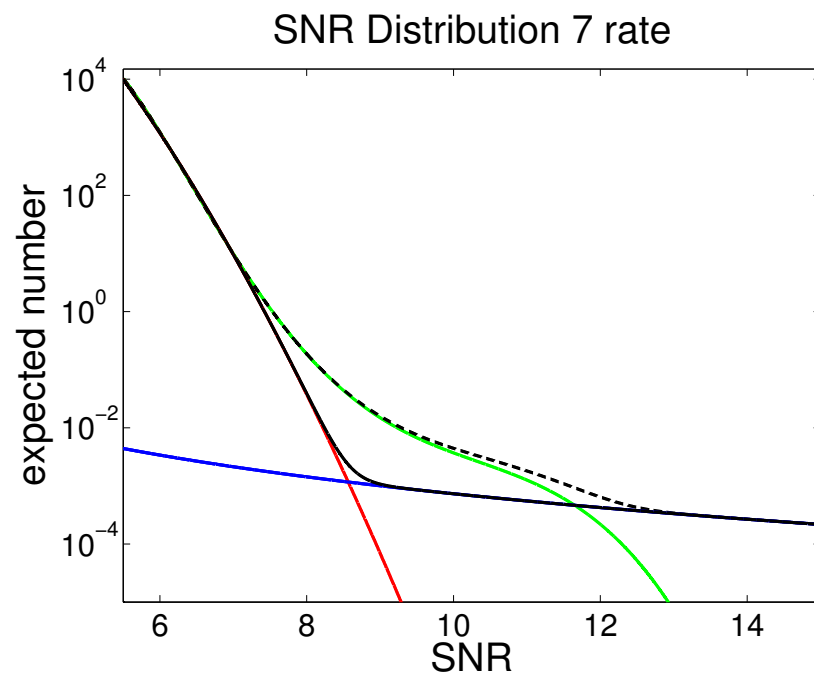


Figure A.6: Reverse CDF of the triggers' SNR for experiment 7: colours assigned as in Fig. A.1.

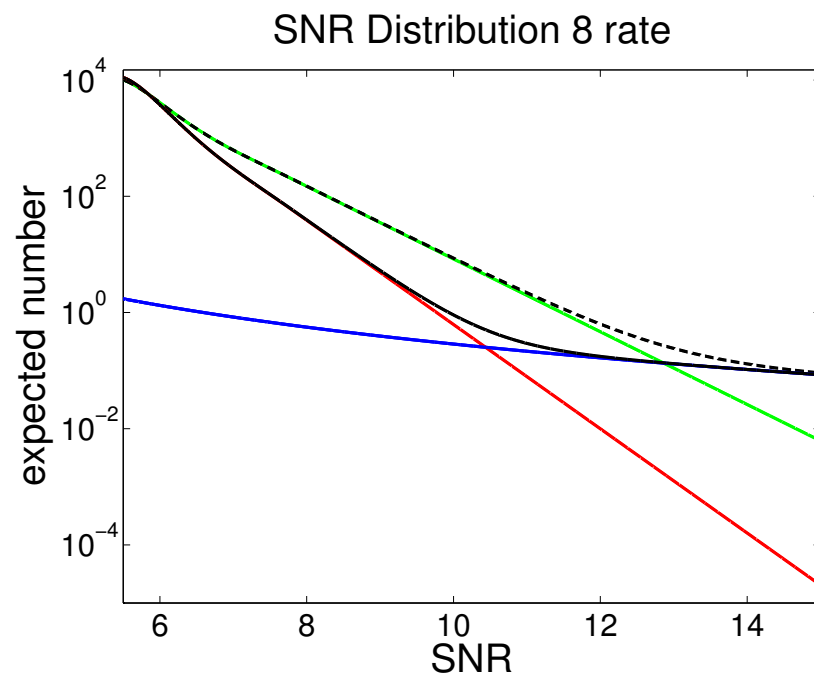


Figure A.7: Reverse CDF of the triggers' SNR for experiment 8: colours assigned as in Fig. A.1.

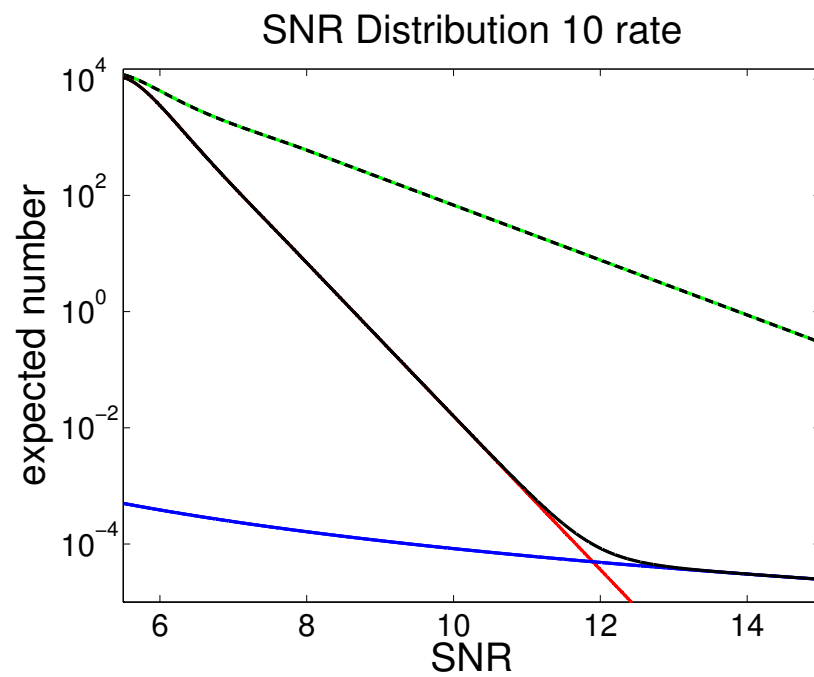


Figure A.8: Reverse CDF of the triggers' SNR for experiment 10: colours assigned as in Fig. A.1.

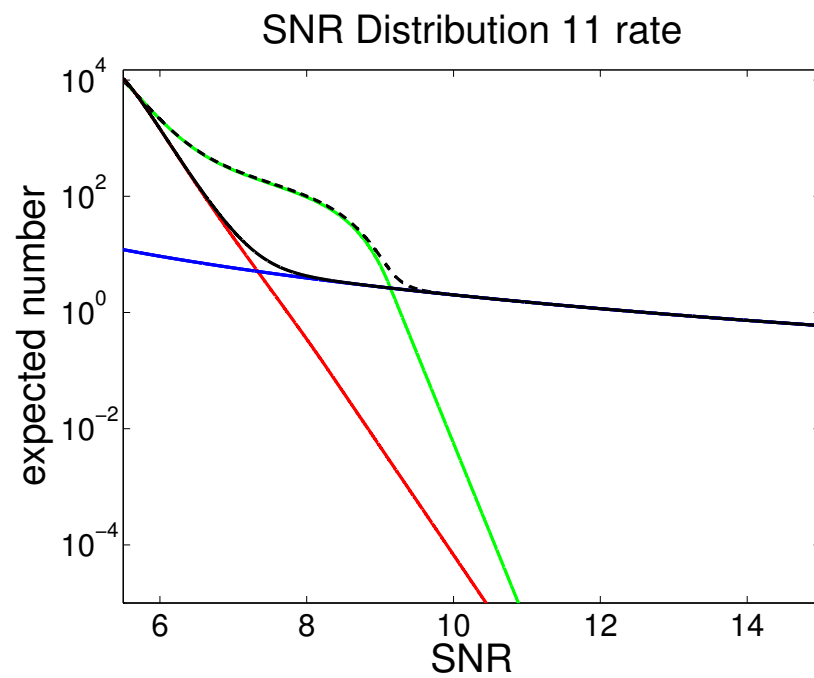


Figure A.9: Reverse CDF of the triggers' SNR for experiment 11: colours assigned as in Fig. A.1.

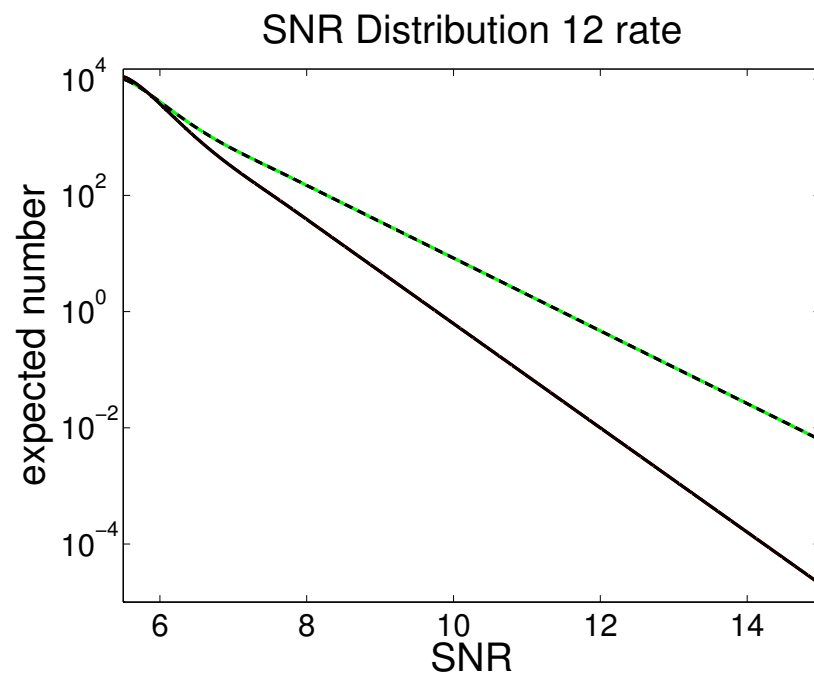


Figure A.10: Reverse CDF of the triggers' SNR for experiment 12: colours assigned as in Fig. A.1.

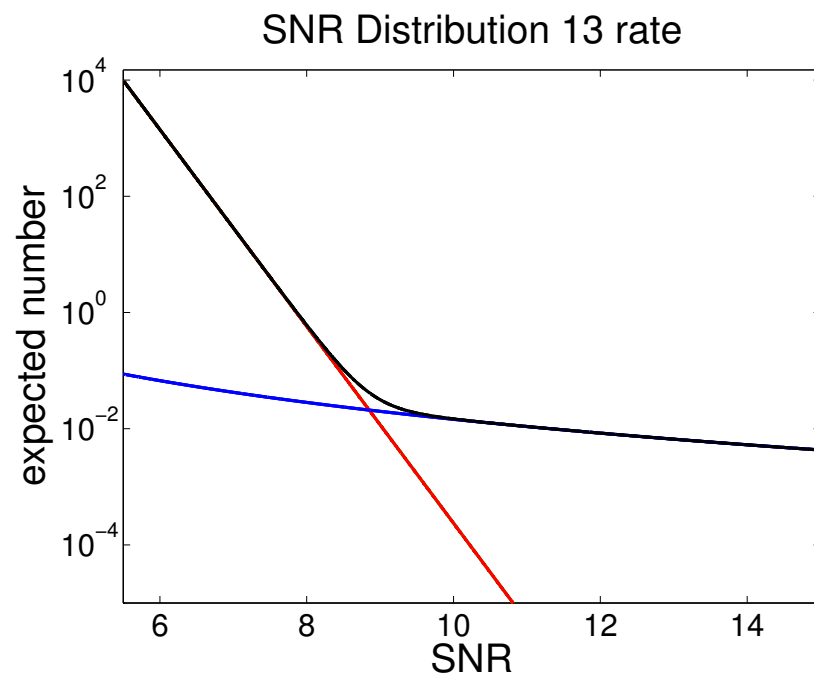


Figure A.11: Reverse CDF of the triggers' SNR for experiment 13: colours assigned as in Fig. A.1.

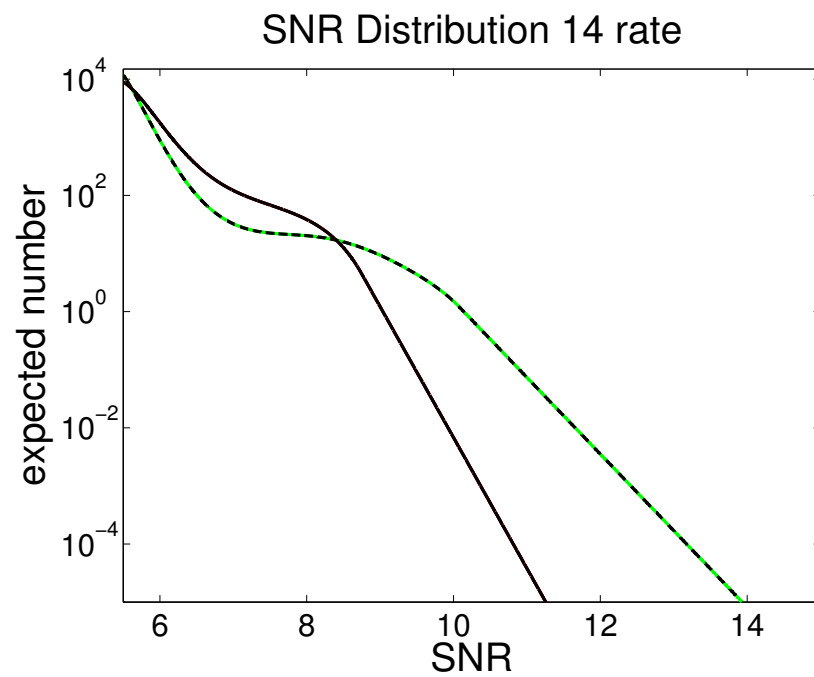
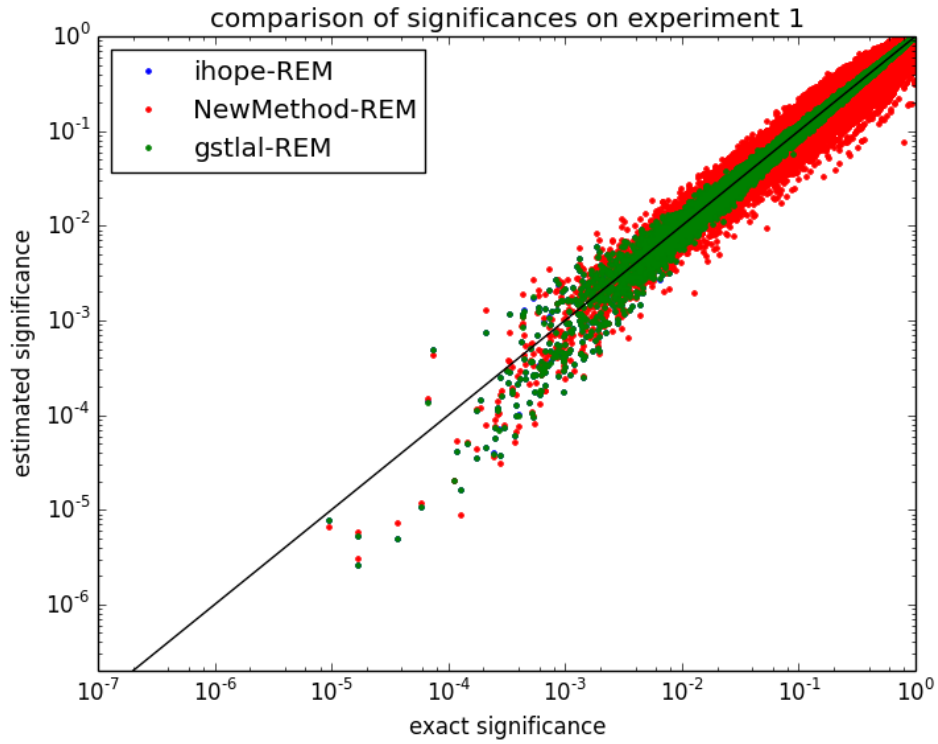


Figure A.12: Reverse CDF of the triggers' SNR for experiment 14: colours assigned as in Fig. A.1.

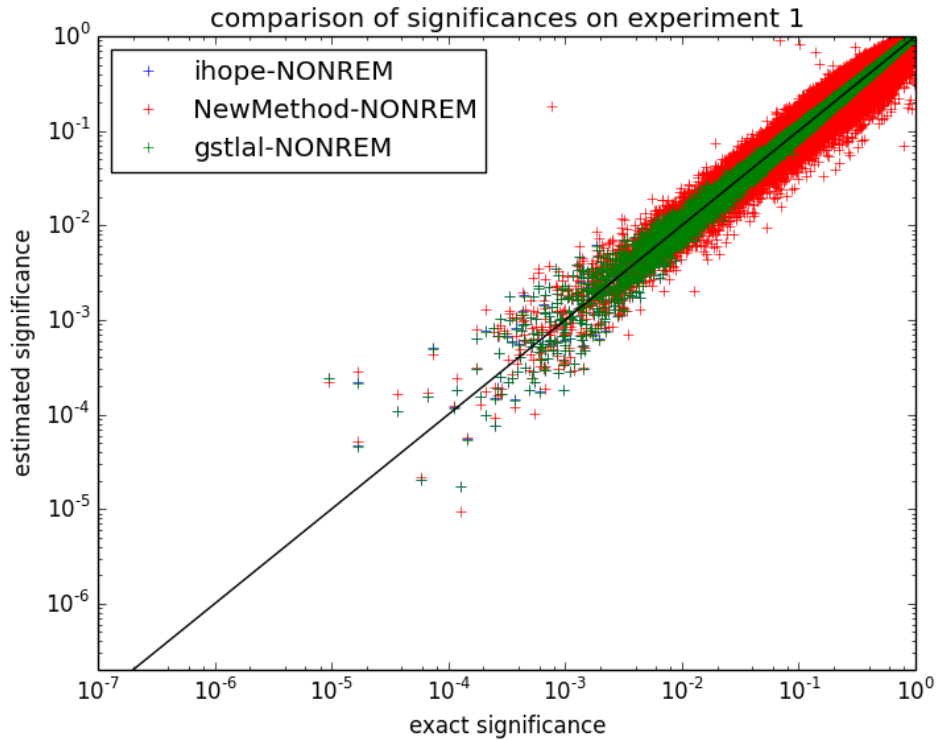


## A.2 Direct comparison

In Figs. A.13, A.14, A.15, A.16, A.17, A.18, A.19, A.20, A.21, and A.22 we present plots of the direct comparison between actual significance and the significance estimations from all methods for each of the experiments.

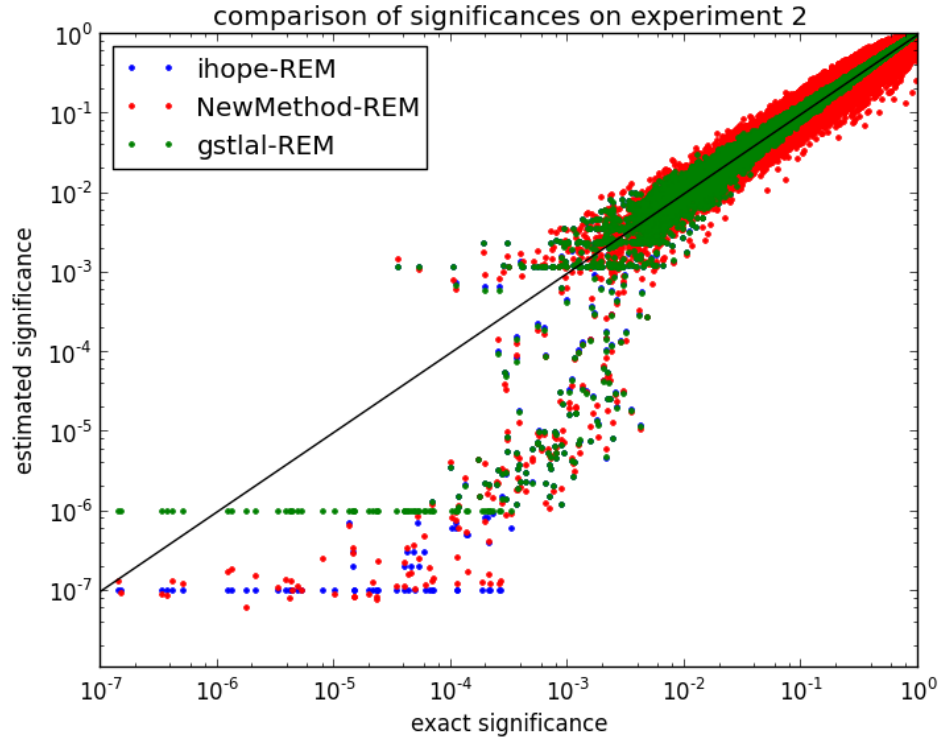


(a) direct comparison with removal

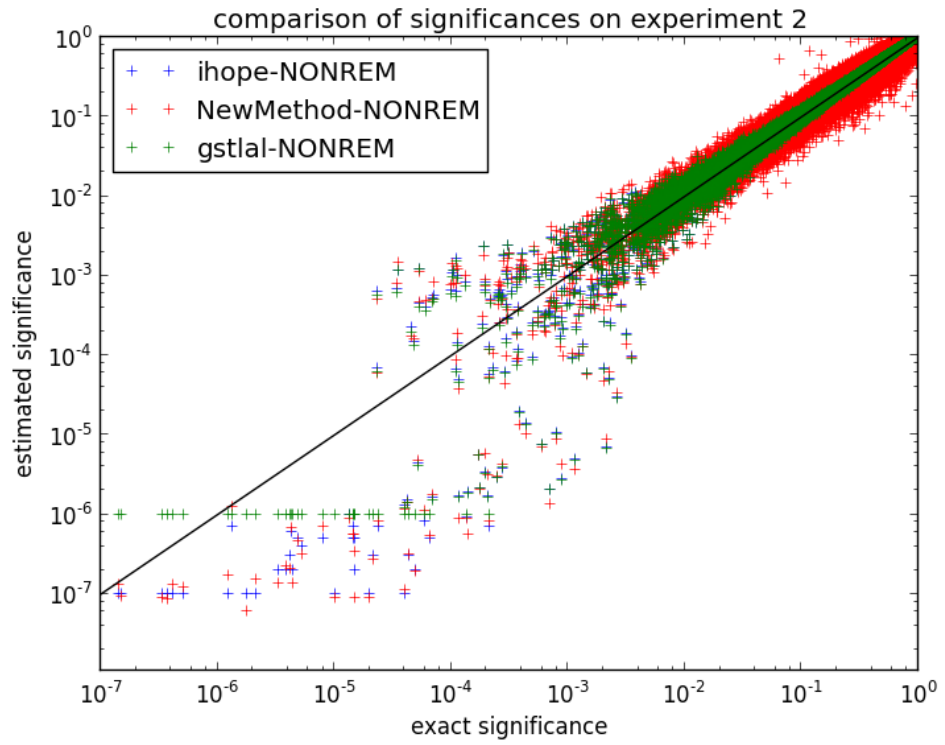


(b) direct comparison with non-removal

Figure A.13: Direct comparisons on experiment 1. Even its background distribution is quite extreme, but since it contains no foreground trigger at all, the estimation is quite concentrated and unbiased. Notice that in remove the points tends to be slightly under the diagonal line, which is being non-conservative.

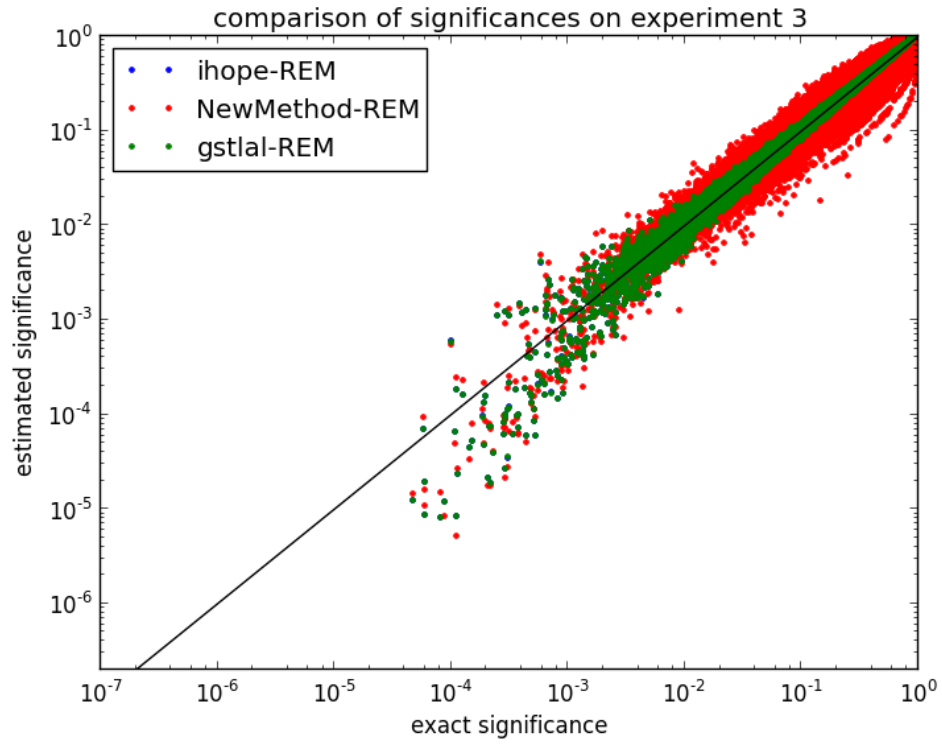


(a) direct comparison with removal

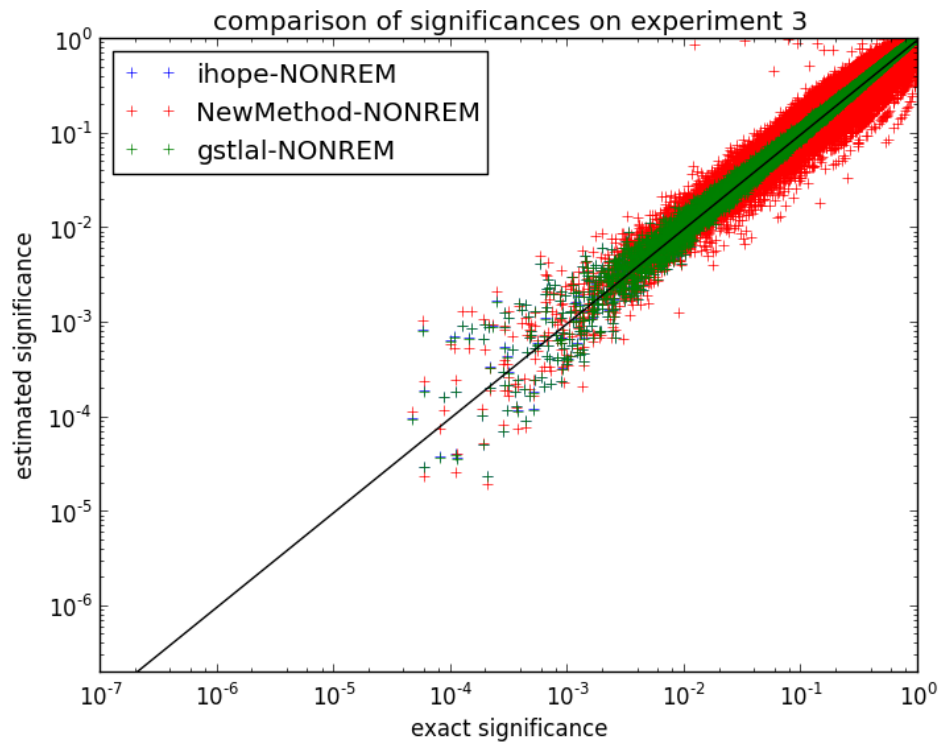


(b) direct comparison with non-removal

Figure A.14: Direct comparisons on experiment 2.

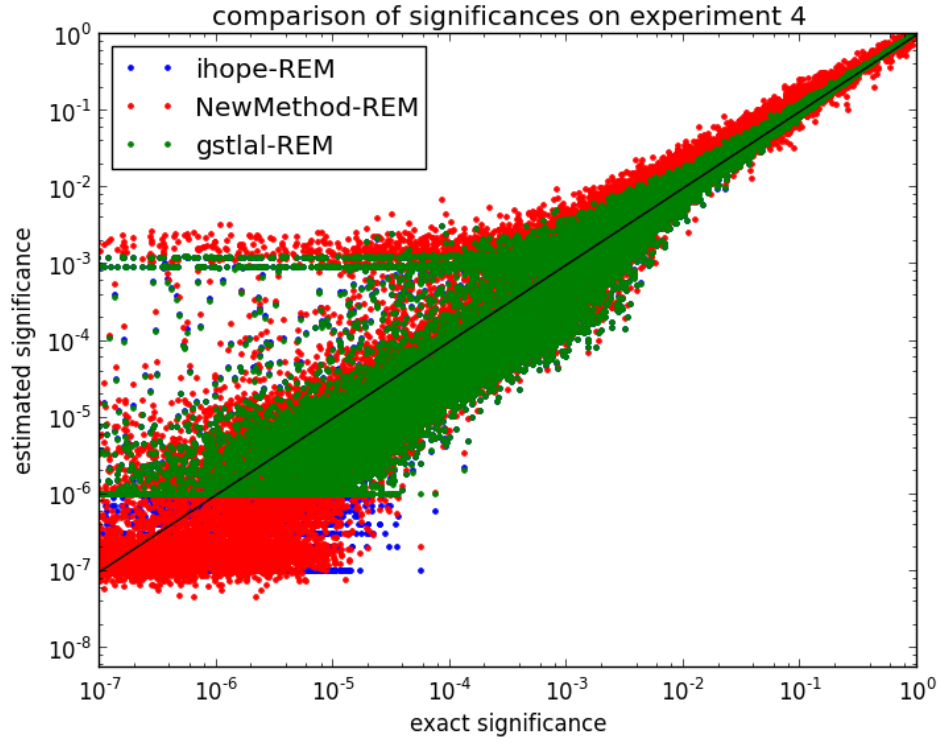


(a) direct comparison with removal

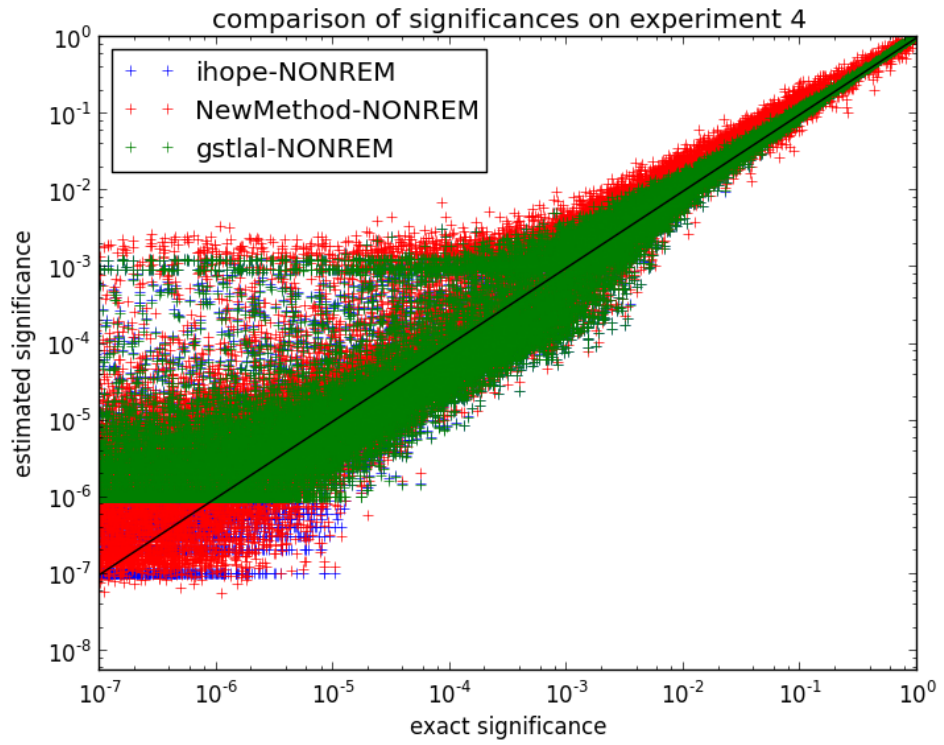


(b) direct comparison with non-removal

Figure A.15: Direct comparisons on experiment 3.

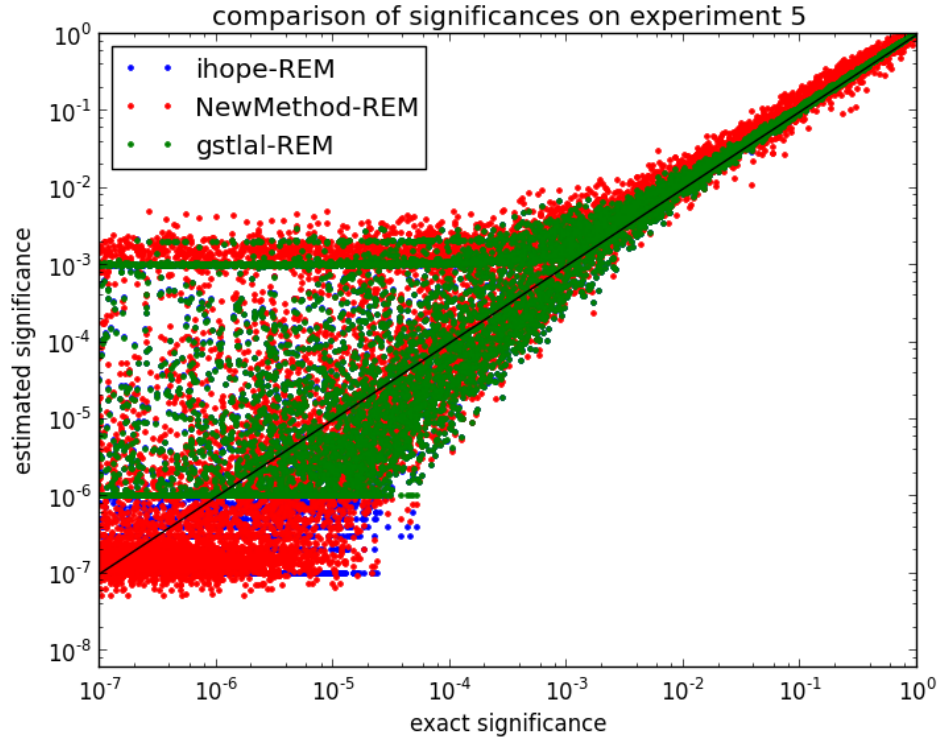


(a) direct comparison with removal

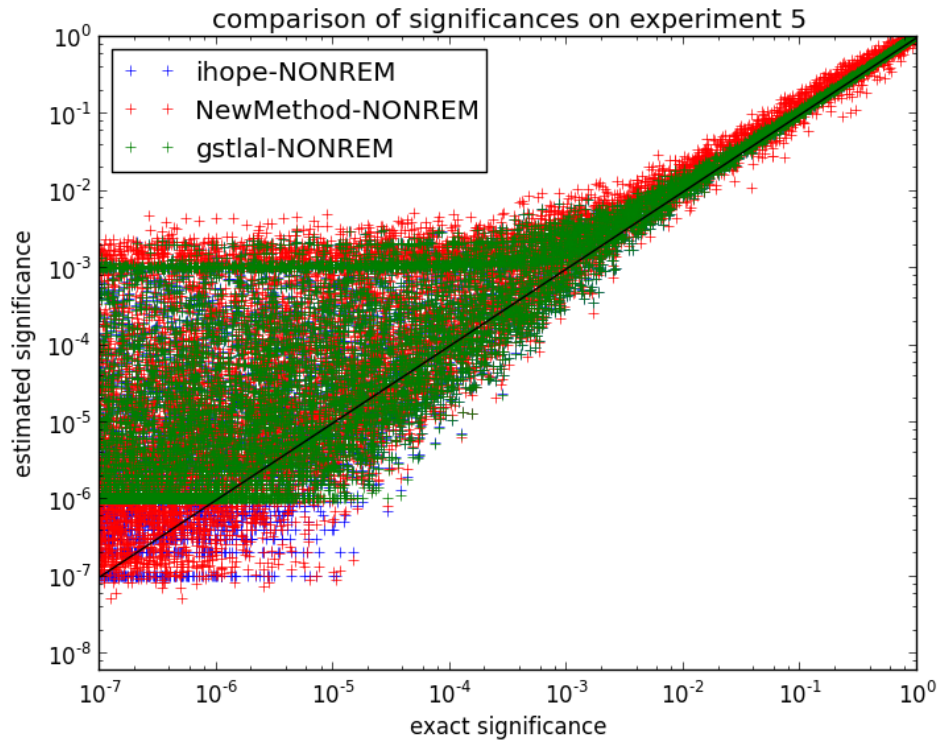


(b) direct comparison with non-removal

Figure A.16: Direct comparisons on experiment 4.

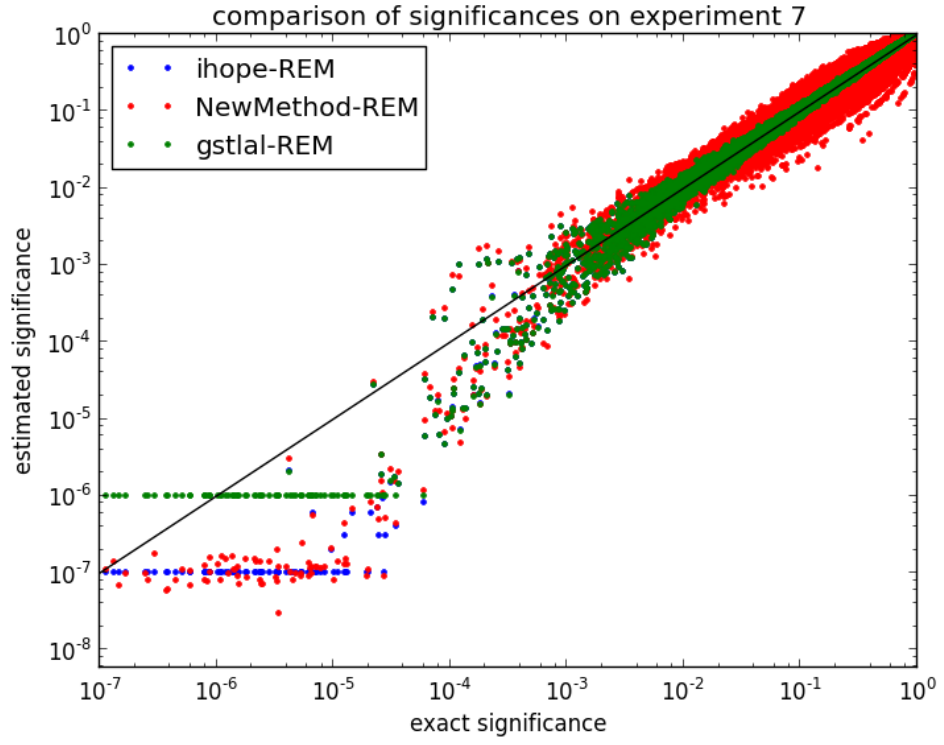


(a) direct comparison with removal

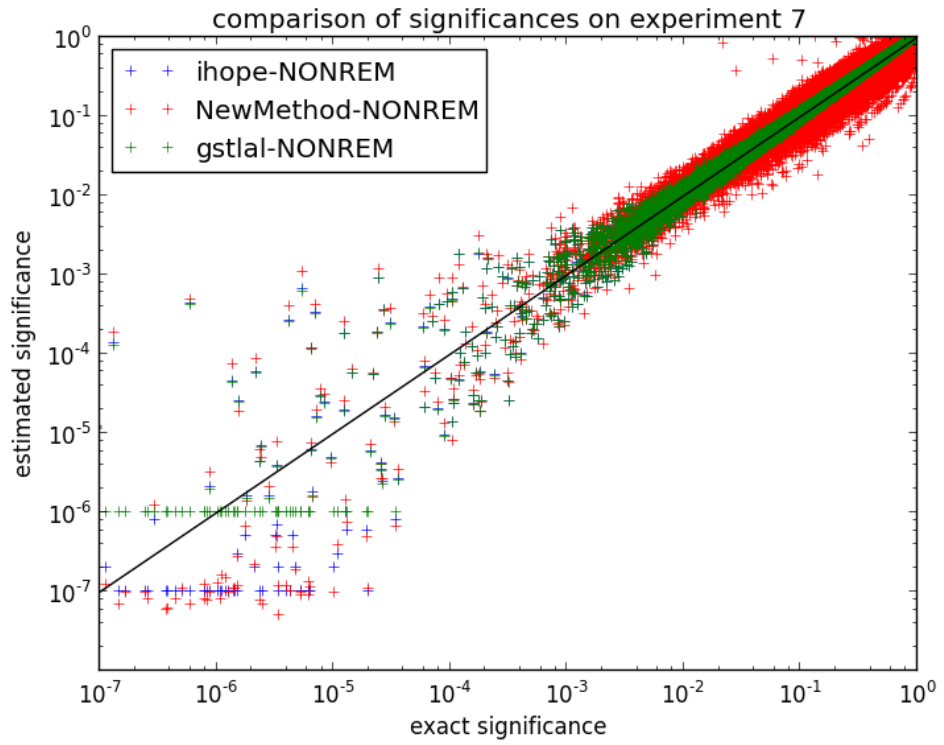


(b) direct comparison with non-removal

Figure A.17: Direct comparisons on experiment 5.

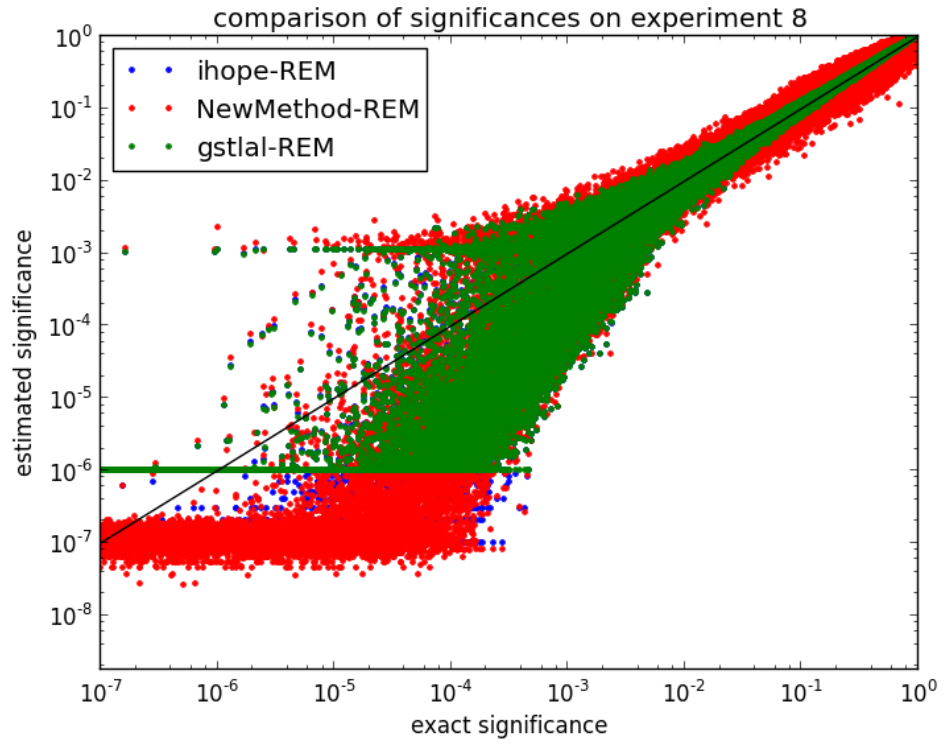


(a) direct comparison with removal

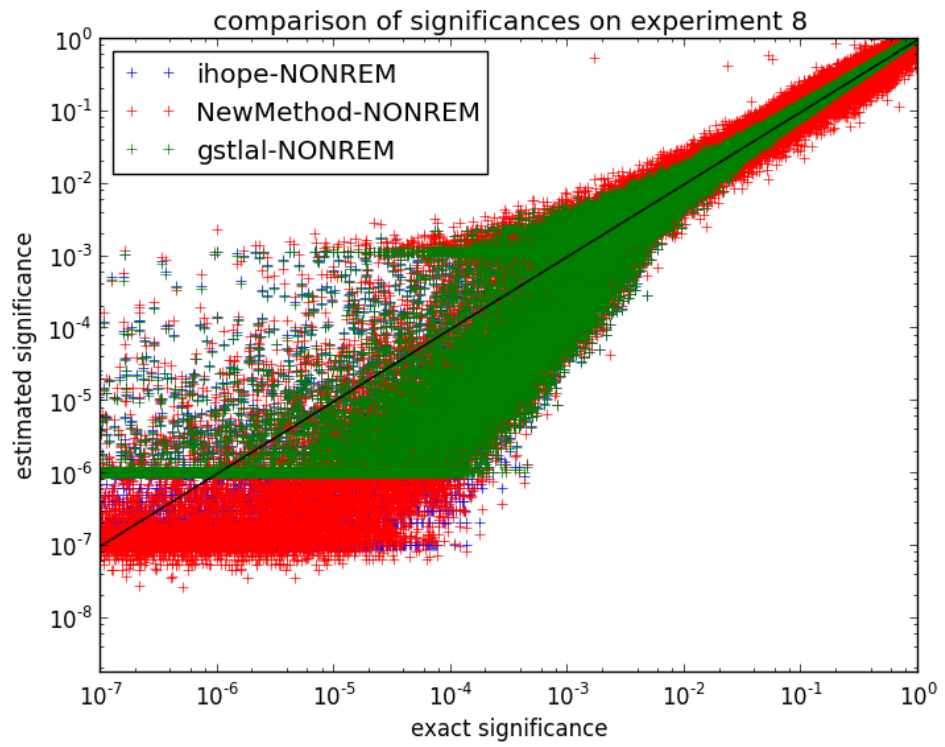


(b) direct comparison with non-removal

Figure A.18: Direct comparisons on experiment 7.



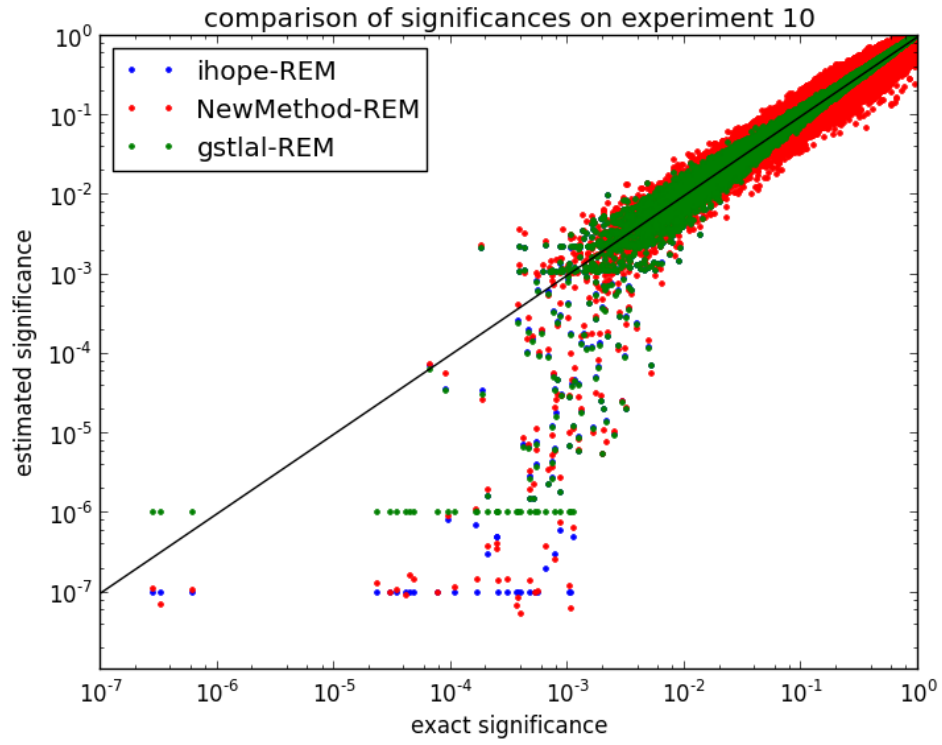
(a) direct comparison with removal



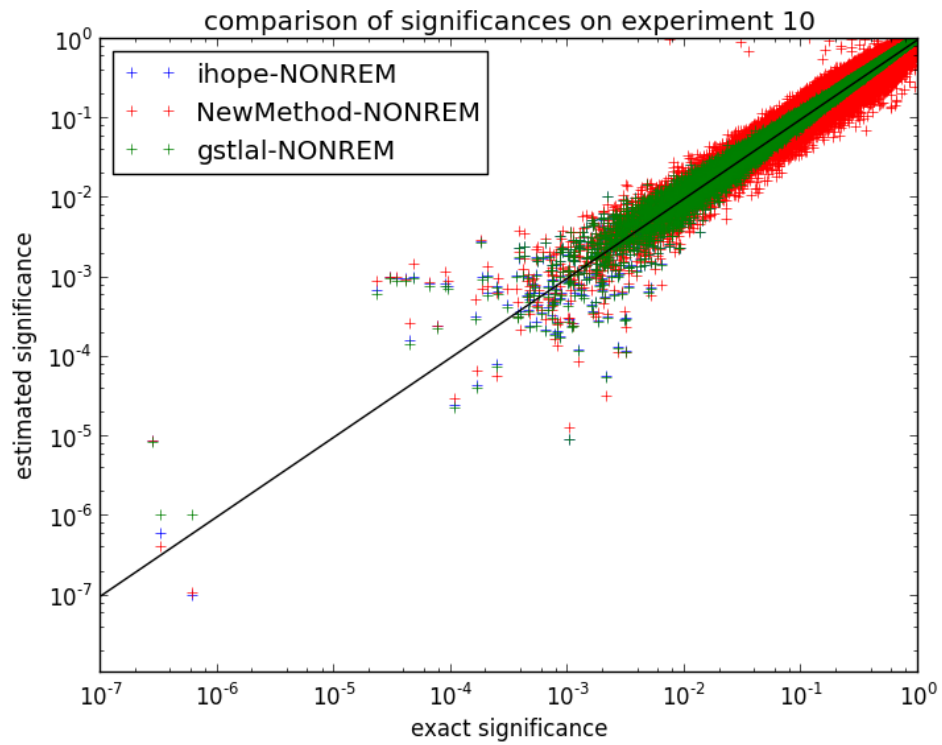
(b) direct comparison with non-removal

Figure A.19: Direct comparisons on experiment 8.



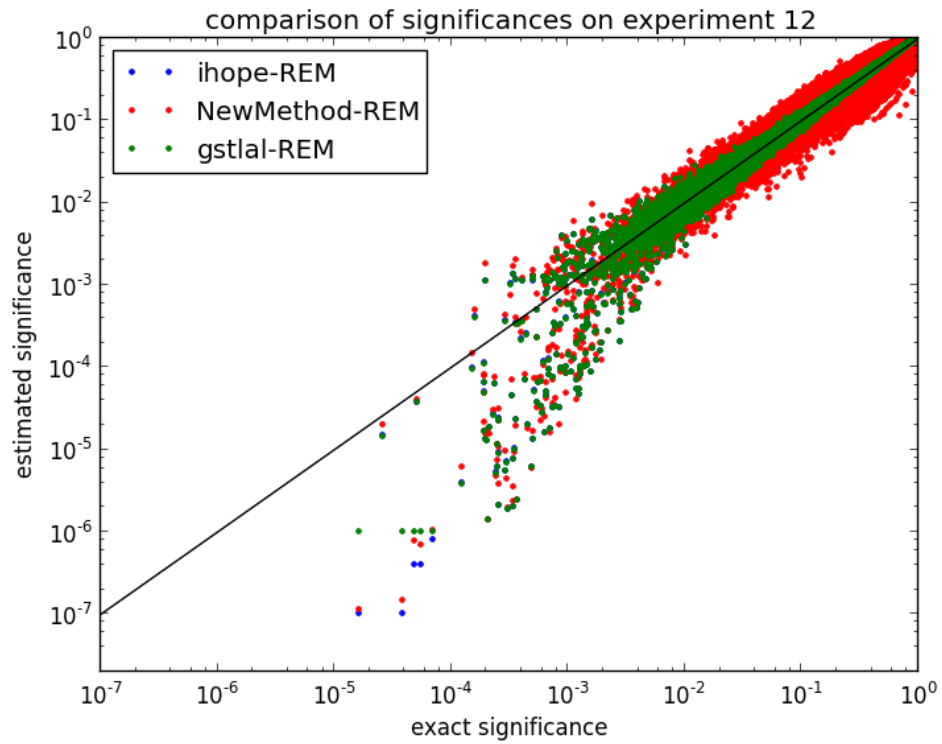


(a) direct comparison with removal

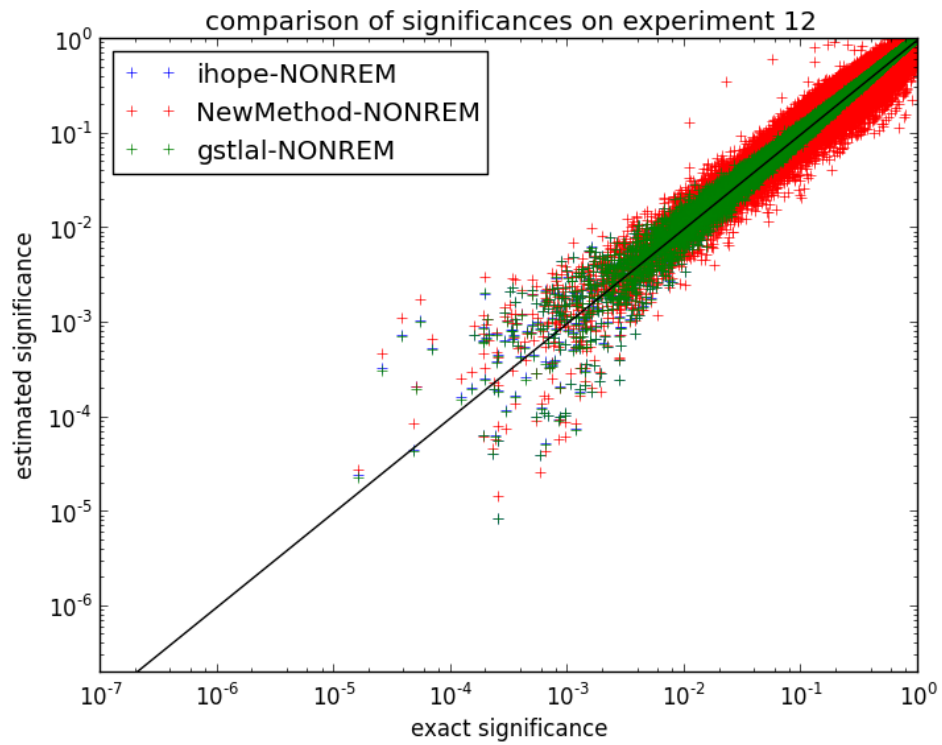


(b) direct comparison with non-removal

Figure A.20: Direct comparisons on experiment 10.

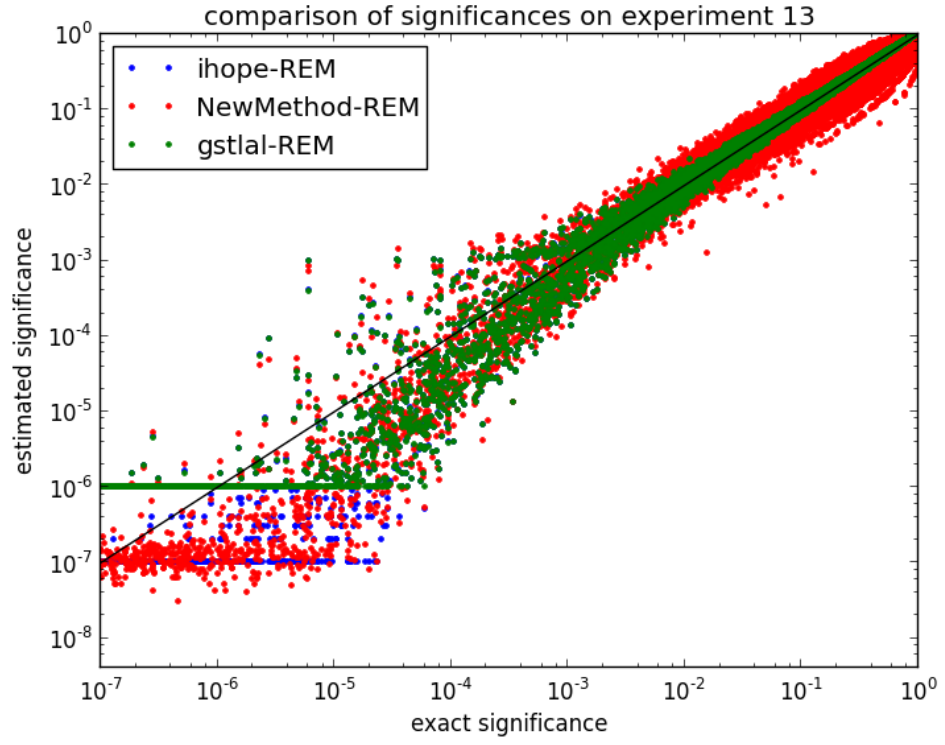


(a) direct comparison with removal

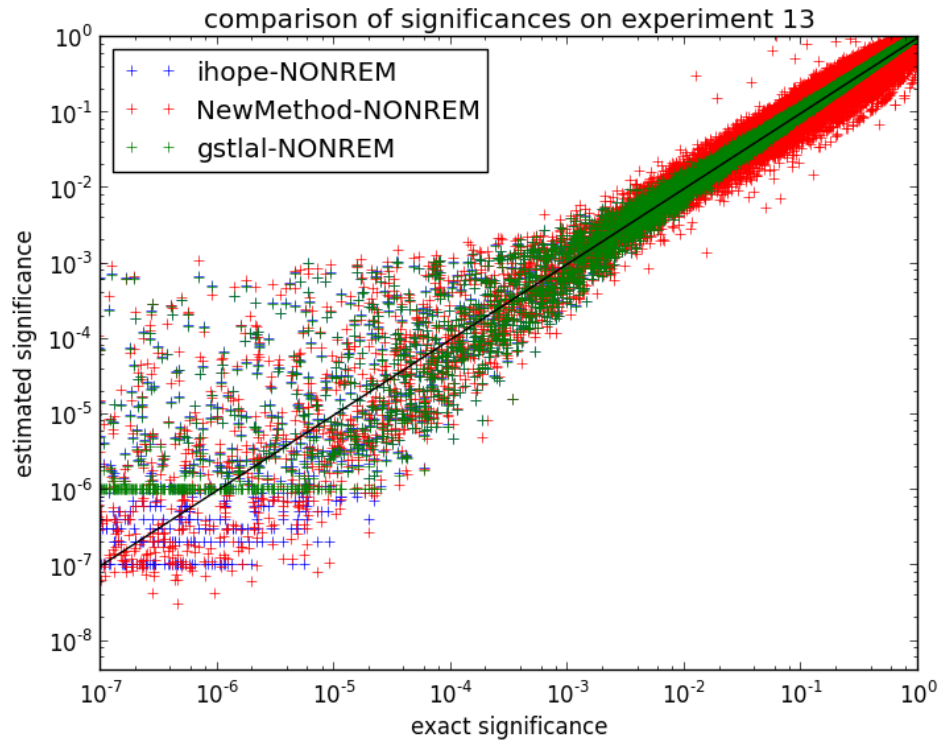


(b) direct comparison with non-removal

Figure A.21: Direct comparisons on experiment 12.



(a) direct comparison with removal



(b) direct comparison with non-removal

Figure A.22: Direct comparisons on experiment 13.

## A.3 Receiver Operating Characteristic plots

In this section we include all remaining ROC plots. Note that it is only possible to create a ROC plot when there are foreground events in the data, so we only show 7 ROC plots, Figs. A.23, A.24, A.25, A.26, A.27, A.28, and A.29, complementing those already shown in Sec. 6.3.3.

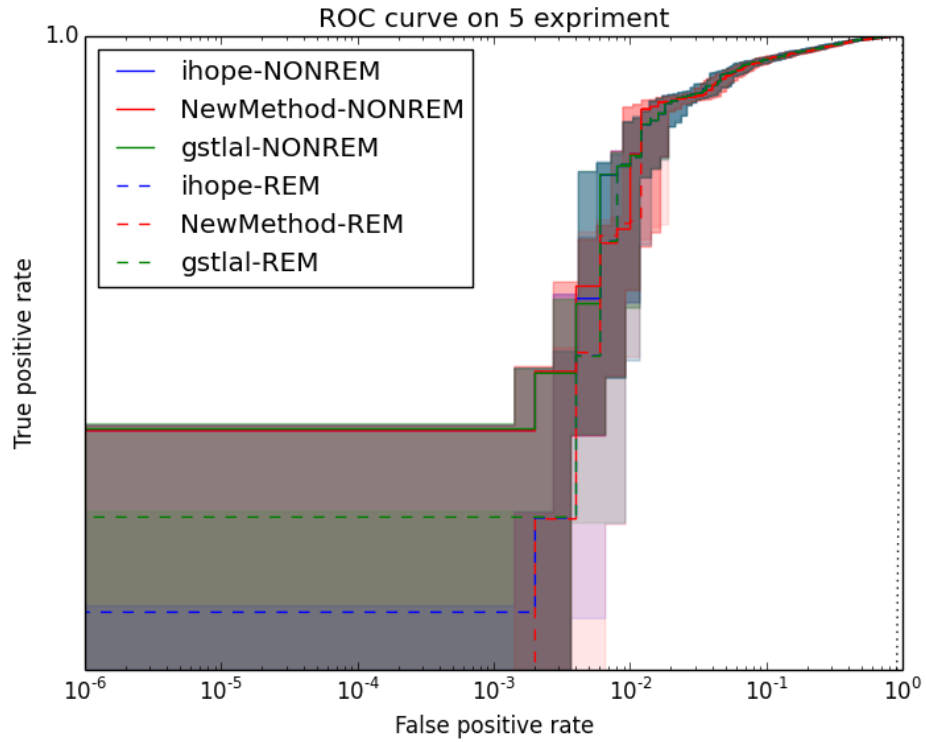


Figure A.23: ROC plot for experiment 5.

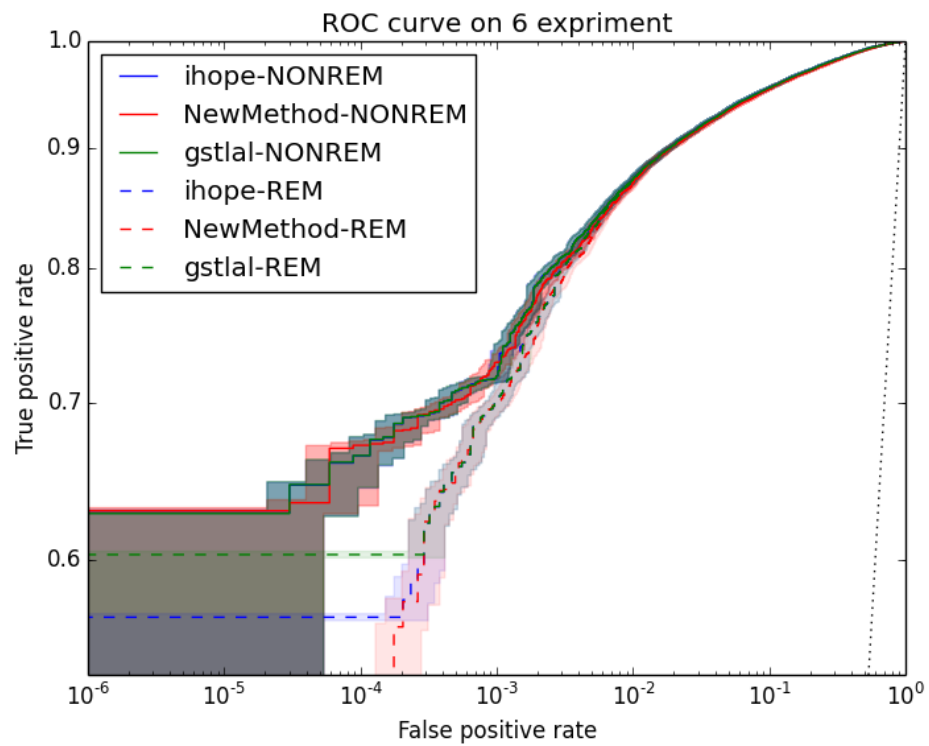


Figure A.24: ROC plot for experiment 6.

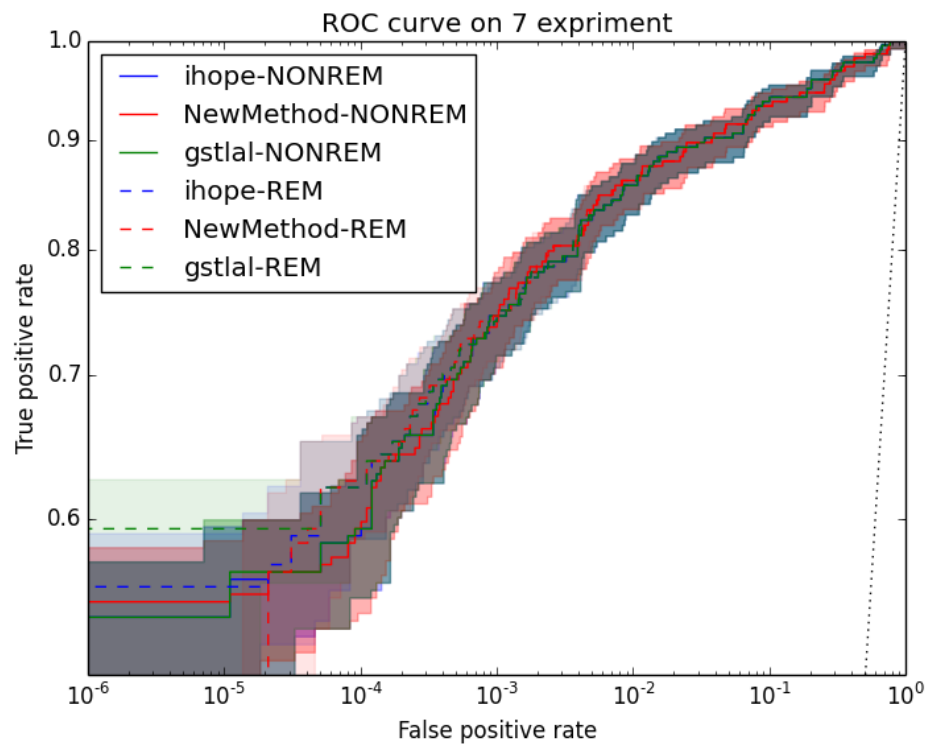


Figure A.25: ROC plot for experiment 7.

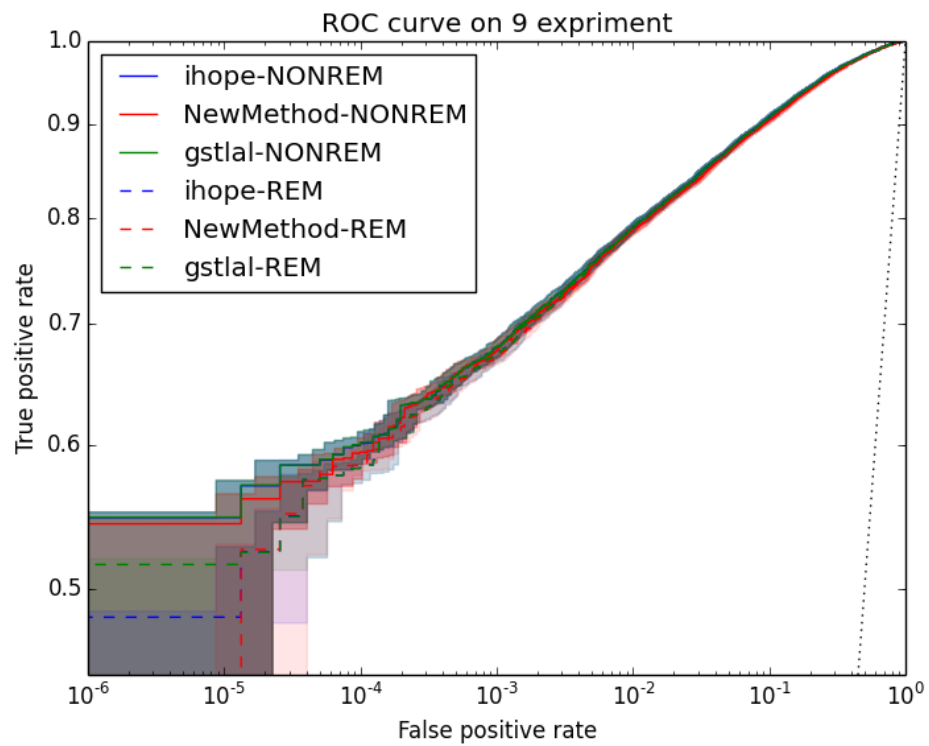


Figure A.26: ROC plot for experiment 9.

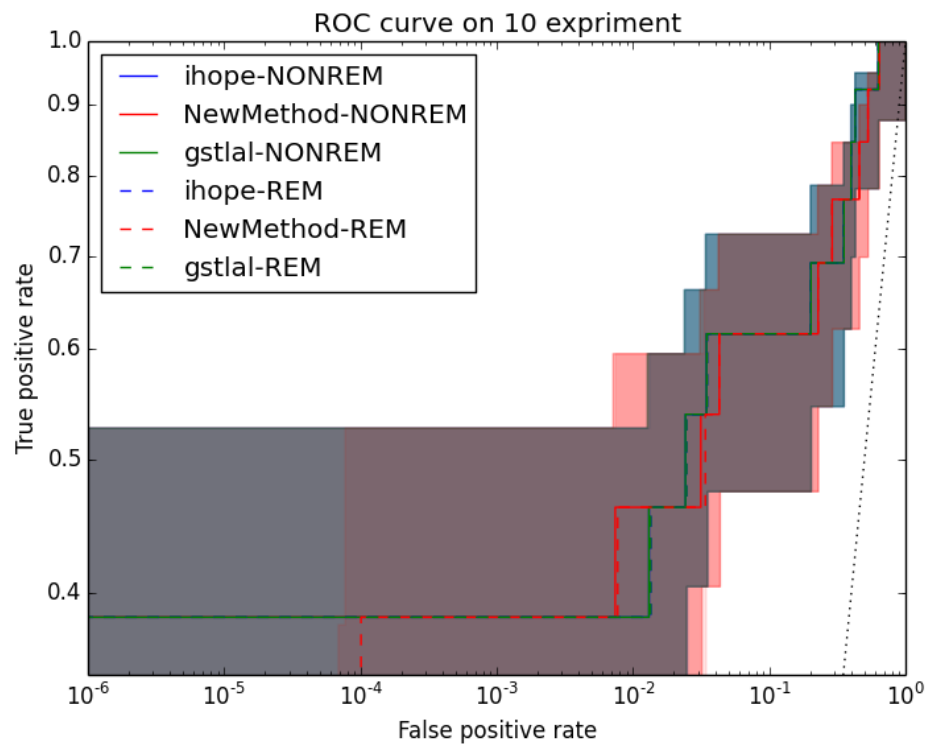


Figure A.27: ROC plot for experiment 10.

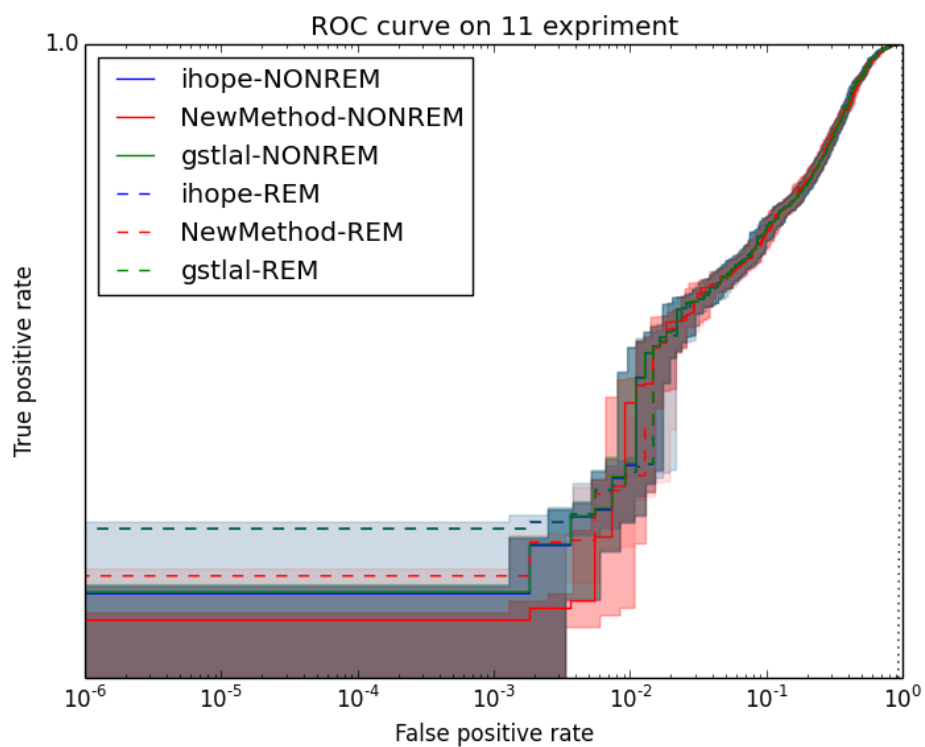


Figure A.28: ROC plot for experiment 11.

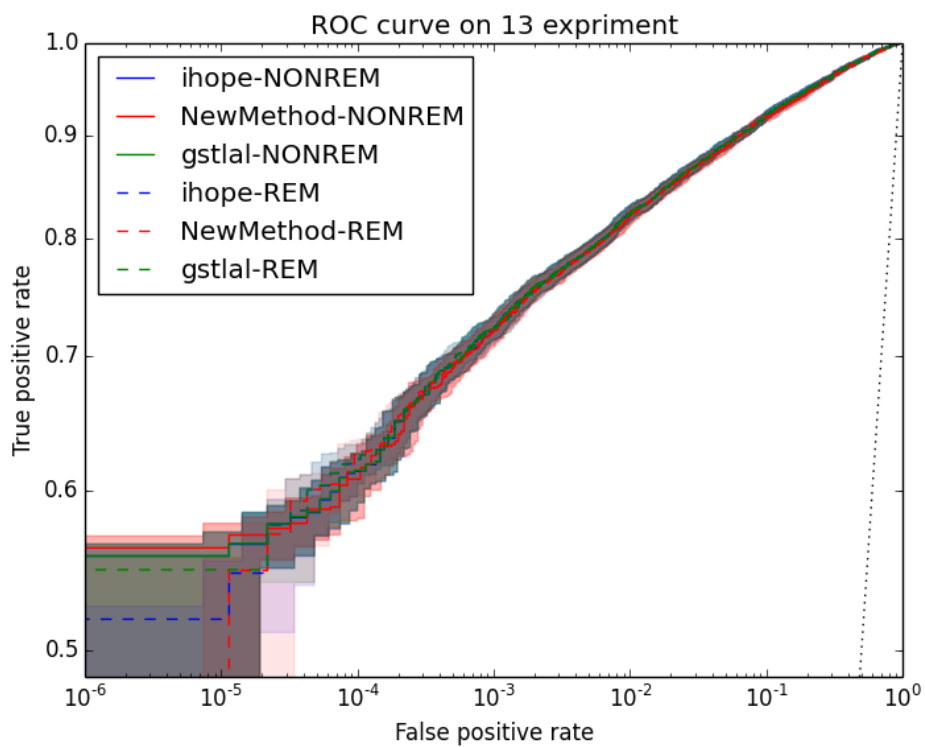


Figure A.29: ROC plot for experiment 13.

## Appendix B

### Parameters for trigger distribution

In this section, we list the parameters used to define the background distributions. Recall that we adopt the form of the SNR distribution for the background triggers according to Eq. 6.1 which rely on the input polynomial coefficients  $a_i$  which are listed in Table B.1 for all 14 experiments.

For the tails of the CDFs, the form is changed to a simpler representation as defined in Eq. 6.1 in order to make sure that the background distribution is well behaved as the SNR rises. The corresponding parameters  $b$ ,  $C_{\text{TP}}$ , and  $\rho_{\text{TP}}$  are listed in Table B.2. Notice that here the actual control parameter is  $C_{\text{TP}}$ , while  $b$  and  $\rho_{\text{TP}}$  are derived values, which could be subject to round-off error.

The rate of both background triggers and foreground triggers are controlled by parameters listed in Table B.3.



Experiment	Background parameters						
	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$
1	-10.0000	-5	0	0	0	0	0
	-10.0000	-5	0	0	0	0	0
2	-7.2240	-0.32	0.53	-0.73	0.12	0.067	-0.018
	-7.2240	-0.32	0.53	-0.73	0.12	0.067	-0.018
3	-7.8000	-3.9	0	0	0	0	0
	-7.8000	-3.9	0	0	0	0	0
4	-6.2800	-0.3	0.5	-0.7	0.1	0.07	-0.02
	-4.8800	-1	0	-0.6	0	-0.04	-0.05
5	-8.4000	-4.2	0	0	0	0	0
	-8.0000	-4	0	0	0	0	0
6	-8.0704	-3	0.8	0.01	-0.05	0.007	-0.0004
	-8.0704	-3	0.8	0.01	-0.05	0.007	-0.0004
7	-9.1072	-4	0.7	0.09	-0.05	0.005	-0.0002
	-9.6200	-5.55	-0.37	0	0	0	0
8	-3.5040	-1.4	0	-0.16	0	-0.034	-0.026
	-4.6400	-2	0	-0.2	0	-0.03	-0.03
9	-7.0000	-3.5	0	0	0	0	0
	-6.6000	-3.3	0	0	0	0	0
10	-2.4800	-1	0	-0.1	0	-0.03	-0.02
	-5.8400	-3	0	-0.1	0	-0.03	-0.03
11	-4.0800	-1	0	-0.3	0	-0.05	-0.03
	-8.3200	-4	0	-0.1	0	-0.035	-0.025
12	-3.5040	-1.4	0	-0.16	0	-0.034	-0.026
	-4.6400	-2	0	-0.2	0	-0.03	-0.03
13	-7.8000	-3.9	0	0	0	0	0
	-7.8000	-3.9	0	0	0	0	0
14	-6.2800	-0.3	0.5	-0.7	0.1	0.07	-0.02
	-4.8800	-1	0	-0.6	0	-0.04	-0.05

Table B.1: parameters for background distribution

Exprmt	IFO	$C_{TP}$	$b$	$\rho_{TP}$
1	1	1e-10	-5.0000	10.1052
1	2	1e-10	-5.0000	10.1052
2	1	1e-4	-1.2690	9.9447
2	2	1e-4	-1.2690	9.9447
3	1	1e-10	-3.9000	11.4041
3	2	1e-10	-3.9000	11.4041
4	1	1e-4	-3.0350	10.0791
4	2	5e-4	-5.1906	8.7474
5	1	1e-10	-4.2000	10.9823
5	2	1e-10	-4.0000	11.2565
6	1	1e-9	-6.8668	13.0625
6	2	1e-9	-6.8668	13.0625
7	1	1e-9	-4.3611	12.9193
7	2	1e-9	-6.8728	9.2876
8	1	2e-2	-1.4415	7.7886
8	2	5e-3	-2.0660	7.8256
9	1	1e-9	-3.5000	11.4209
9	2	1e-9	-3.3000	11.7798
10	1	5e-2	-1.0889	8.0018
10	2	1e-3	-3.0410	7.8544
11	1	3e-4	-7.1603	9.1251
11	2	1e-5	-4.2931	8.2823
12	1	2e-2	-1.4415	7.7886
12	2	5e-3	-2.0660	7.8256
13	1	1e-9	-3.9000	10.8137
13	2	1e-9	-3.9000	10.8137
14	1	1e-4	-3.0350	10.0791
14	2	5e-4	-5.1906	8.7474

Table B.2: Parameters for background distribution (tail). Note that only  $C_{TP}$  is the controlled parameter and so it's accurate, while  $b$  and  $\rho_{TP}$  are derived values and subject to numerical error.

Exprmt	$\lambda_1$	$\lambda_2$	$n$	measured $n$	AstroRate (loud)	AstroRate (all)
1	10500	10500	11.025	11.0098	0	0
1_withSignal	10500	10500	11.025	11.0154	3.95	8.55
2	11500	11500	13.225	13.2453	0.001	0.0022
3	9900	9900	9.801	9.7874	0	0
4	12000	9000	10.8	10.8023	2.96	6.41
5	9800	10100	9.898	9.8857	2.74	5.94
6	8000	15000	12	12.0195	0.548	1.19
7	10300	9900	10.197	10.2206	0.0011	0.0024
8	10100	11100	11.211	11.2202	0.438	0.95
9	9700	10600	10.282	10.2785	0.11	0.24
10	12000	10800	12.96	12.9552	0.0001	0.0003
11	9800	10700	10.486	10.4938	3.07	6.65
12	10100	11100	11.211	11.2047	0	0
13	9900	9900	9.801	9.7814	0.022	0.048
14	12000	9000	10.8	10.7857	0	0

Table B.3: parameters for rates of both background and foreground signals

# Bibliography

- [1] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, Misner, C. W., Thorne, K. S., & Wheeler, J. A., Ed. W. H. Freeman, 1973.
- [2] B. S. Sathyaprakash and B. F. Schutz, “Physics, Astrophysics and Cosmology with Gravitational Waves,” *Living Rev. Relativity*, p. 137, Mar. 2009. [Online]. Available: <http://www.emis.ams.org/journals/LRG/Articles/lrr-2009-2/download/lrr-2009-2BW.pdf>
- [3] J. D. E. Creighton and W. G. Anderson, *Gravitational-wave physics and astronomy. An introduction to theory, experiment and data analysis*. Hoboken, NJ: John Wiley & Sons, 2011.
- [4] S. W. Hawking, W. Israel, and W. G. Unruh, “300 Years of Gravitation,” *Physics Today*, vol. 41, no. 7, p. 69, 1988. [Online]. Available: <http://link.aip.org/link/PHTOAD/v41/i7/p69/s1&Agg=doi>
- [5] J. H. Taylor and J. M. Weisberg, “A new test of general relativity - Gravitational radiation and the binary pulsar PSR 1913+16,” *Astrophysical Journal*, vol. 253, pp. 908–920, 1982. [Online]. Available: [http://adsabs.harvard.edu/cgi-bin/nph-data\\_query?bibcode=1982ApJ...253..908T&link\\_type=ABSTRACT&delimiter=026E30F\\$npapers2://publication/doi/10.1086/159690](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1982ApJ...253..908T&link_type=ABSTRACT&delimiter=026E30F$npapers2://publication/doi/10.1086/159690)
- [6] J. M. Weisberg and J. H. Taylor, “Relativistic Binary Pulsar B1913+16: Thirty Years of Observations and Analysis,” *Astronomical Society of the Pacific Conference Series*, vol. 328, p. 25, 2005. [Online]. Available: <http://arxiv.org/abs/astro-ph/0407149>
- [7] R. a. Hulse and J. H. Taylor, “A High-Sensitivity Pulsar Survey,” *The Astrophysical Journal*, vol. 191, p. L59, 1974.
- [8] R. A. Hulse and J. H. Taylor, “Discovery of a pulsar in a binary system,” *The Astrophysical Journal Letters*, vol. 195, pp. L51–L53, Jan. 1975.

- [9] M. Kramer, I. H. Stairs, R. N. Manchester, M. a. McLaughlin, a. G. Lyne, R. D. Ferdman, M. Burgay, D. R. Lorimer, a. Possenti, N. D'Amico, J. M. Sarkissian, G. B. Hobbs, J. E. Reynolds, P. C. C. Freire, and F. Camilo, "Tests of general relativity from timing the double pulsar." *Science*, vol. 314, no. October, pp. 97–102, 2006.
- [10] P. a. R. Ade, R. W. Aikin, D. Barkats, S. J. Benton, C. a. Bischoff, J. J. Bock, J. a. Brevik, I. Buder, E. Bullock, C. D. Dowell, L. Duband, J. P. Filippini, S. Fliescher, S. R. Golwala, M. Halpern, M. Hasselfield, S. R. Hildebrandt, G. C. Hilton, V. V. Hristov, K. D. Irwin, K. S. Karkare, J. P. Kaufman, B. G. Keating, S. a. Kernasovskiy, J. M. Kovac, C. L. Kuo, E. M. Leitch, M. Lueker, P. Mason, C. B. Netterfield, H. T. Nguyen, R. O'Brient, R. W. Ogburn IV, a. Orlando, C. Pryke, C. D. Reintsema, S. Richter, R. Schwarz, C. D. Sheehy, Z. K. Staniszewski, R. V. Sudiwala, G. P. Teply, J. E. Tolan, a. D. Turner, a. G. Vieregg, C. L. Wong, and K. W. Yoon, "Detection of B - mode polarization at degree angular scales by BICEP2," *Physical Review Letters*, vol. 112, pp. 1–26, 2014.
- [11] E. Berti, E. Barausse *et al.*, "Testing General Relativity with Present and Future Astrophysical Observations," *arXiv preprint arXiv:1501.07274*, p. 186, 2015. [Online]. Available: <http://arxiv.org/abs/1501.07274>
- [12] J. Logue, C. D. Ott, I. S. Heng, P. Kalmus, and J. H. C. Scargill, "Inferring core-collapse supernova physics with gravitational waves," *Physical Review D*, vol. 86, no. 4, p. 044023, Aug. 2012. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.86.044023>
- [13] P. A. Evans, J. P. Osborne, J. A. Kennea, M. Smith, D. M. Palmer, N. Gehrels, J. M. Gelbord, A. Homeier, M. Voge, N. L. Strotjohann, D. F. Cowen, S. Boser, M. Kowalski, and A. Stasik, "Swift follow-up of IceCube triggers, and implications for the Advanced-LIGO era," *Monthly Notices of the Royal Astronomical Society*, vol. 448, no. 3, pp. 2210–2223, Jan. 2015. [Online]. Available: <http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/stv136>
- [14] S. Rosswog, "The multi-messenger picture of compact binary mergers," *arXiv preprint arXiv:1501.02081*, p. 40, Dec. 2015. [Online]. Available: <http://arxiv.org/abs/1501.02081>
- [15] B. F. Schutz, "Determining the Hubble constant from gravitational wave observations," pp. 310–311, 1986.
- [16] L. S. Finn, "Detection, Measurement and Gravitational Radiation," *Physical Review D*, vol. 46, no. 12, p. 37, Sep. 1992. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.46.5236>

- [17] C. Cutler and E. Flanagan, “Gravitational waves from merging compact binaries: How accurately can one extract the binarys parameters from the inspiral waveform?” *Physical Review D*, vol. 49, no. 6, pp. 2658–2697, Mar. 1994. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.49.2658>
- [18] B. Schutz, *A First Course in General Relativity*. Cambridge University Press, May 2009.
- [19] Gravitational wave-related images. <http://www.johnstonsarchive.net/relativity/pictures.html>.
- [20] L.-X. Li and B. Paczyński, “Transient Events from Neutron Star Mergers,” *The Astrophysical Journal*, vol. 507, no. 1, pp. L59–L62, 1998. [Online]. Available: <http://arxiv.org/abs/astro-ph/9807272>
- [21] E. Berger, W. Fong, and R. Chornock, “AN r -PROCESS KILONOVA ASSOCIATED WITH THE SHORT-HARD GRB 130603B,” *The Astrophysical Journal*, vol. 774, p. L23, Jun. 2013. [Online]. Available: <http://iopscience.iop.org/2041-8205/774/2/L23>
- [22] C. Messenger and J. Read, “Measuring a cosmological distance-redshift relationship using only gravitational wave observations of binary neutron star coalescences,” *Physical Review Letters*, vol. 108, pp. 15–19, 2012.
- [23] B. D. Metzger and E. Berger, “What is the Most Promising Electromagnetic Counterpart of a Neutron Star Binary Merger?” *The Astrophysical Journal*, vol. 746, no. 1, p. 16, Aug. 2011. [Online]. Available: <http://stacks.iop.org/0004-637X/746/i=1/a=48?key=crossref.d68f6fb5c3cb50c996482a71c321c4d5>
- [24] L. Blanchet, “Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries,” *Living Reviews in Relativity*, vol. 9, 2006. [Online]. Available: <http://www.livingreviews.org/lrr-2006-4>
- [25] G. Faye, S. Marsat, L. Blanchet, and B. R. Iyer, “The third and a half-post-Newtonian gravitational wave quadrupole mode for quasi-circular inspiralling compact binaries,” *Classical and Quantum Gravity*, vol. 29, no. 17, p. 175004, Sep. 2012. [Online]. Available: <http://stacks.iop.org/0264-9381/29/i=17/a=175004?key=crossref.ac0a6ccf441b4a87b6b0a5a8f3c12ea4>
- [26] J. Aasi *et al.*, “The NINJA-2 project: detecting and characterizing gravitational waveforms modelled using numerical binary black hole simulations,” *Classical and Quantum Gravity*, vol. 31, no. 11, p. 115004, Jun. 2014. [Online]. Available: <http://iopscience.iop.org/0264-9381/31/11/115004>

- [27] T. W. Baumgarte and S. L. Shapiro, *Numerical Relativity: Solving Einstein's Equations on the Computer*. Cambridge University Press, 2010.
- [28] Y. Sekiguchi, K. Kiuchi, K. Kyutoku, and M. Shibata, "Gravitational Waves and Neutrino Emission from the Merger of Binary Neutron Stars," *Physical Review Letters*, vol. 107, no. 5, p. 051102, Jul. 2011. [Online]. Available: <http://arxiv.org/abs/1105.2125><http://link.aps.org/doi/10.1103/PhysRevLett.107.051102>
- [29] C. Kalaghatgi, P. Ajith, and K. G. Arun, "Template-space metric for searches for gravitational waves from the inspiral, merger and ringdown of binary black holes," *arXiv preprint arXiv: 1501.04418*, p. 8, Jan. 2015. [Online]. Available: <http://arxiv.org/abs/1501.4418>
- [30] E. Barausse, A. Buonanno, S. A. Hughes, G. Khanna, S. OSullivan, and Y. Pan, "Modeling multipolar gravitational-wave emission from small mass-ratio mergers," *Physical Review D*, vol. 85, no. 2, p. 024046, Jan. 2012. [Online]. Available: <http://prd.aps.org/abstract/PRD/v85/i2/e024046>
- [31] J. S. Read, L. Baiotti, J. D. E. Creighton, J. L. Friedman, B. Giacomazzo, K. Kyutoku, C. Markakis, L. Rezzolla, M. Shibata, and K. Taniguchi, "Matter effects on binary neutron star waveforms," *Physical Review D*, vol. 88, no. 4, p. 044042, Aug. 2013. [Online]. Available: <http://journals.aps.org/prd/abstract/10.1103/PhysRevD.88.044042>
- [32] A. Bauswein and H. Janka, "Measuring Neutron-Star Properties via Gravitational Waves from Neutron-Star Mergers," *Physical Review Letters*, vol. 108, no. 1, p. 011101, Jan. 2012. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevLett.108.011101>
- [33] Y. Pan, A. Buonanno, A. Taracchini, M. Boyle, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, B. Szilágyi, and A. Zenginoglu, "Stability of nonspinning effective-one-body model in approximating two-body dynamics and gravitational-wave emission," *Physical Review D - Particles, Fields, Gravitation and Cosmology*, vol. 89, no. 6, p. 061501, 2014.
- [34] R. J. E. Smith, K. Cannon, C. Hanna, D. Keppel, and I. Mandel, "Towards rapid parameter estimation on gravitational waves from compact binaries using interpolated waveforms," *Physical Review D*, vol. 87, no. 12, p. 122002, Jun. 2013. [Online]. Available: <http://journals.aps.org/prd/abstract/10.1103/PhysRevD.87.122002>
- [35] P. Canizares, S. E. Field, J. R. Gair, and M. Tiglio, "Gravitational wave parameter estimation with compressed likelihood evaluations," *Physical Review D - Particles, Fields, Gravitation and Cosmology*, vol. 87, p. 124005, 2013.

- [36] M. Pürrer, “Frequency-domain reduced order models for gravitational waves from aligned-spin compact binaries,” *Classical and Quantum Gravity*, vol. 31, no. 19, p. 195010, Oct. 2014. [Online]. Available: <http://stacks.iop.org/0264-9381/31/i=19/a=195010?key=crossref.8a044bd1683c82deba7e1da8cd1e6f36>
- [37] S. Nissanke, D. E. Holz, S. A. Hughes, N. Dalal, and J. L. Sievers, “EXPLORING SHORT GAMMA-RAY BURSTS AS GRAVITATIONAL-WAVE STANDARD SIRENS,” *The Astrophysical Journal*, vol. 725, no. 1, pp. 496–514, Dec. 2010. [Online]. Available: <http://iopscience.iop.org/0004-637X/725/1/496>
- [38] C. J. Moore, R. H. Cole, and C. P. L. Berry, “Gravitational-wave sensitivity curves,” *Classical and Quantum Gravity*, vol. 32, no. 1, p. 015014, Jan. 2015. [Online]. Available: <http://stacks.iop.org/0264-9381/32/i=1/a=015014?key=crossref.6d77a04a033873eb201d87a385da7158>
- [39] K. Postnov and L. Yungelson, “Close Binary White Dwarfs, Neutron Stars, Black Holes: Formation, Evolution, and Related Phenomena,” *arXiv preprint arXiv:1403.4754*, p. 181, Mar. 2014. [Online]. Available: <http://arxiv.org/abs/1403.4754>
- [40] Gravitational wave observatory. [http://en.wikipedia.org/wiki/Gravitational-wave\\_observatory](http://en.wikipedia.org/wiki/Gravitational-wave_observatory).
- [41] J. Veitch, V. Raymond, B. Farr, and W. Farr, “Robust parameter estimation for compact binaries with ground-based gravitational-wave observations using LALInference,” *Physical Review D*, in press. [Online]. Available: <http://arxiv.org/abs/1409.7215>
- [42] J. Weber, “Gravitational radiation,” *Physical Review Letters*, vol. 18, no. 13, pp. 498–501, 1967.
- [43] —, “Physical review letters 16,” *Physical Review Letters*, vol. 22, no. 24, pp. 1320–1324, 1969. [Online]. Available: [http://prl.aps.org/pdf/PRL/v22/i24/p1320\\_1](http://prl.aps.org/pdf/PRL/v22/i24/p1320_1)
- [44] —, “Anisotropy and polarization in the gravitational-radiation experiments,” *Physical Review Letters*, vol. 25, no. 3, pp. 180–184, 1970.
- [45] J. Hough, “The search for gravitational waves,” in *AIP Conference Proceedings*, vol. 957, 2007, pp. 93–98.
- [46] R. W. P. DREVER, J. HOUGH, R. BLAND, and G. W. LESSNOFF, “Search for Short Bursts of Gravitational Radiation,” pp. 340–344, 1973.



- [47] J. Aasi *et al.*, “Parameter estimation for compact binary coalescence signals with the first generation gravitational-wave detector network,” *Physical Review D*, vol. 88, no. 6, p. 062001, Sep. 2013. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.88.062001>
- [48] Nobel prize official web site. [www.nobelprize.org/nobel\\_prizes/physics/laureates/1993/press.html](http://www.nobelprize.org/nobel_prizes/physics/laureates/1993/press.html).
- [49] O. D. Aguiar, “Past, present and future of the Resonant-Mass gravitational wave detectors,” *Research in Astronomy and Astrophysics*, vol. 11, no. 1, pp. 1–42, 2010. [Online]. Available: <http://arxiv.org/abs/1009.1138>
- [50] Auriga sensitivity. [www.auriga.lnl.infn.it/auriga/detector/run1/sensitivity.html](http://www.auriga.lnl.infn.it/auriga/detector/run1/sensitivity.html).
- [51] B. P. Abbott *et al.*, “Ligo: the laser interferometer gravitational-wave observatory,” *Reports on Progress in Physics*, vol. 72, no. 7, p. 076901, 2009. [Online]. Available: <http://stacks.iop.org/0034-4885/72/i=7/a=076901>
- [52] T. Accadia *et al.*, “Virgo: a laser interferometer to detect gravitational waves,” *Journal of Instrumentation*, vol. 7, no. 03, p. P03012, 2012. [Online]. Available: <http://stacks.iop.org/1748-0221/7/i=03/a=P03012>
- [53] The LIGO Scientific Collaboration, “Beating the spin-down limit on gravitational wave emission from the Vela pulsar,” *The Astronomical Journal*, vol. 737, no. 2, p. 93, 2011. [Online]. Available: <http://arxiv.org/abs/1104.2712>
- [54] —, “Beating the spin-down limit on gravitational wave emission from the Vela pulsar,” *The Astrophysical Journal Letters*, vol. 683, no. 1, pp. 45–49, 2008. [Online]. Available: <http://dx.doi.org/10.1086/591526><http://arxiv.org/abs/1104.2712>
- [55] J. E. Staff, P. Jaikumar, V. Chan, and R. Ouyed, “SPINDOWN OF ISOLATED NEUTRON STARS: GRAVITATIONAL WAVES OR MAGNETIC BRAKING?” *The Astrophysical Journal*, vol. 751, no. 1, p. 24, May 2012. [Online]. Available: <http://iopscience.iop.org/0004-637X/751/1/24>
- [56] G. M. Harry and the LIGO Scientific Collaboration, “Advanced ligo: the next generation of gravitational wave detectors,” *Classical and Quantum Gravity*, vol. 27, no. 8, p. 084006, 2010. [Online]. Available: <http://stacks.iop.org/0264-9381/27/i=8/a=084006>
- [57] F. Acernese *et al.*, “Advanced Virgo: a 2nd generation interferometric gravitational wave detector,” *Classical and Quantum Gravity*, vol. 32, p. 024001, 2015.

- [58] M. Punturo *et al.*, “The third generation of gravitational wave observatories and their science reach,” *Classical and Quantum Gravity*, vol. 27, no. 8, p. 084007, 2010. [Online]. Available: <http://stacks.iop.org/0264-9381/27/i=8/a=084007>
- [59] Comparison of sensitive volume for initial and advanced ligo. <https://dcc.ligo.org/LIGO-D0901491/public>.
- [60] B. Iyer, T. Souradeep, C. Unnikrishnan, S. Dhurandhar, S. Raja, and A. Sengupta. (2011) Ligo-india technical report. [Online]. Available: <https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=75988>
- [61] K. Somiya, “Detector configuration of kagrathe japanese cryogenic gravitational-wave detector,” *Classical and Quantum Gravity*, vol. 29, no. 12, p. 124007, 2012. [Online]. Available: <http://stacks.iop.org/0264-9381/29/i=12/a=124007>
- [62] H. Grote, “The GEO 600 status,” p. 084003, 2010.
- [63] S. Hild, “Beyond the second generation of laser-interferometric gravitational wave observatories,” p. 124006, 2012.
- [64] S. Hild *et al.*, “Sensitivity studies for third-generation gravitational wave observatories,” *Classical and Quantum Gravity*, vol. 28, no. 9, p. 094013, 2011. [Online]. Available: <http://stacks.iop.org/0264-9381/28/i=9/a=094013>
- [65] The Isc instrument science white paper for 2014-2015. <https://dcc.ligo.org/LIGO-T1400316/public>.
- [66] Lisa. <http://lisa.nasa.gov/>.
- [67] P. Amaro-Seoane, S. Aoudia, S. Babak, P. Binétruy, E. Berti, A. Bohé, C. Caprini, M. Colpi, N. J. Cornish, K. Danzmann, J.-f. Dufaux, J. Gair, O. Jennrich, P. Jetzer, A. Klein, R. N. Lang, A. Lobo, T. Littenberg, S. T. McWilliams, G. Nelemans, A. Petiteau, E. K. Porter, B. F. Schutz, A. Sesana, R. Stebbins, T. Sumner, M. Vallisneri, S. Vitale, M. Volonteri, and H. Ward, “eLISA: Astrophysics and cosmology in the millihertz regime,” *GW Notes*, vol. 6, pp. 4–110, 2012. [Online]. Available: <http://arxiv.org/abs/1201.3621>
- [68] Lisa pathfinder. [http://www.esa.int/Our\\_Activities/Space\\_Science/LISA\\_Pathfinder\\_overview](http://www.esa.int/Our_Activities/Space_Science/LISA_Pathfinder_overview).
- [69] S. Kawamura *et al.*, “The Japanese space gravitational wave antenna - DECIGO,” p. 012006, 2008.

- [70] G. M. Harry, P. Fritschel, D. a. Shaddock, W. Folkner, and E. S. Phinney, “Laser interferometry for the Big Bang Observer,” pp. 7361–7361, 2006.
- [71] K. J. Lee, N. Wex, M. Kramer, B. W. Stappers, C. G. Bassa, G. H. Janssen, R. Karuppusamy, and R. Smits, “Gravitational wave astronomy of single sources with a pulsar timing array,” *Monthly Notices of the Royal Astronomical Society*, vol. 414, pp. 3251–3264, 2011.
- [72] R. M. L. J. Baker, G. V. Stephensonb, and F. Li, “Analyses of the Frequency and Intensity of Laboratory Generated HFGWs,” in *AIP Conference Proceeding 969*, 2008, p. 1036. [Online]. Available: <http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.2844940>
- [73] S. Aoyama, R. Tazai, and K. Ichiki, “Upper limit on the amplitude of gravitational waves around 0.1 Hz from the Global Positioning System BRIEF REPORTS BRIEF REPORTS,” *Physical Review D - Particles, Fields, Gravitation and Cosmology*, vol. 89, p. 067101, Feb. 2014. [Online]. Available: <http://arxiv.org/abs/1402.4521>
- [74] *Lecture from Jan Harms.*
- [75] L. P. Singer, L. R. Price, B. Farr, A. L. Urban, C. Pankow, S. Vitale, J. Veitch, W. M. Farr, C. Hanna, K. Cannon, T. Downes, P. Graff, C.-J. Haster, I. Mandel, T. Sidery, and A. Vecchio, “THE FIRST TWO YEARS OF ELECTROMAGNETIC FOLLOW-UP WITH ADVANCED LIGO AND VIRGO,” *The Astrophysical Journal*, vol. 795, no. 2, p. 105, Oct. 2014. [Online]. Available: <http://stacks.iop.org/0004-637X/795/i=2/a=105?key=crossref.6c5af1d41b1de3e8c5e5d65e7f126e0b>
- [76] C. P. L. Berry, I. Mandel, H. Middleton, L. P. Singer, A. L. Urban, A. Vecchio, S. Vitale, K. Cannon, B. Farr, W. M. Farr, P. B. Graff, C. Hanna, C.-J. Haster, S. Mohapatra, C. Pankow, L. R. Price, T. Sidery, and J. Veitch, “Parameter estimation for binary neutron-star coalescences with realistic noise during the Advanced LIGO era,” *arXiv preprint arXiv: 1411.6934*, p. 20, Nov. 2014. [Online]. Available: <http://arxiv.org/abs/1411.6934>
- [77] K. Cannon, C. Hanna, and D. Keppel, “Method to estimate the significance of coincident gravitational-wave observations from compact binary coalescence,” *Physical Review D - Particles, Fields, Gravitation and Cosmology*, vol. 88, no. 2, p. 024025, Jul. 2013. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.88.024025>
- [78] M. Ws, M.-A. Bizouard, V. Brisson, F. Cavalier, M. Davier, P. Hello, N. Leroy, F. Robinet, and M. Vavoulidis, “On the background estimation by time slides in a

- network of gravitational wave detectors,” *Classical and Quantum Gravity*, vol. 27, no. 1, p. 015005, Jan. 2010. [Online]. Available: <http://stacks.iop.org/0264-9381/27/i=1/a=015005?key=crossref.4d9b07e8dfb51c254307744862e5a725>
- [79] J. Abadie *et al.*, “Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors,” *Classical and Quantum Gravity*, vol. 27, no. 17, p. 173001, Sep. 2010. [Online]. Available: <http://stacks.iop.org/0264-9381/27/i=17/a=173001?key=crossref.8cd38004165a7603a0d966b5b7b8f15d>
- [80] J. Clark, L. Cadonati, J. Healy, I. S. Heng, J. Logue, N. Mangini, L. London, L. Pekowsky, and D. Shoemaker, “Gravitational Wave Astrophysics Using Pulsars,” *Astrophysics and Space Science Proceedings*, vol. 40, pp. 281–287, Jun. 2015. [Online]. Available: <http://arxiv.org/abs/1406.5426>
- [81] P. A. Evans, J. P. Osborne, J. A. Kennea, M. Smith, D. M. Palmer, N. Gehrels, J. M. Gelbord, A. Homeier, M. Voge, N. L. Strotjohann, D. F. Cowen, S. Boeser, M. Kowalski, and A. Stasik, “Swift follow-up of IceCube triggers, and implications for the Advanced-LIGO era,” *arXiv preprint arXiv: 1501.04435*, vol. 18, no. January, p. 18, Jan. 2015. [Online]. Available: <http://arxiv.org/abs/1501.4435>
- [82] I. Mandel, L. Z. Kelley, and E. Ramirez-Ruiz, “Towards Improving the Prospects for Coordinated Gravitational-Wave and Electromagnetic Observations,” *Proceedings of the International Astronomical Union*, vol. 7, no. S285, pp. 358–360, Apr. 2012. [Online]. Available: [http://www.journals.cambridge.org/abstract\\_S1743921312001068](http://www.journals.cambridge.org/abstract_S1743921312001068)
- [83] D. M. Coward, E. J. Howell, T. Piran, G. Stratta, M. Branchesi, O. Bromberg, B. Gendre, R. R. Burman, and D. Guetta, “The Swift short gamma-ray burst rate density: implications for binary neutron star merger rates,” *Monthly Notices of the Royal Astronomical Society*, vol. 425, no. 4, pp. 2668–2673, Oct. 2012. [Online]. Available: <http://mnras.oxfordjournals.org/cgi/doi/10.1111/j.1365-2966.2012.21604.x>
- [84] C. Palenzuela, L. Lehner, M. Ponce, S. L. Liebling, M. Anderson, D. Neilsen, and P. Motl, “Electromagnetic and gravitational outputs from binary-neutron-star coalescence,” *Physical Review Letters*, vol. 111, p. 6, Jan. 2013. [Online]. Available: <http://arxiv.org/abs/1301.7074>
- [85] B. D. Metzger, A. Bauswein, S. Goriely, and D. Kasen, “Neutron-powered precursors of kilonovae,” *Mon. Not. R. Astr. Soc.*, no. September, p. 6, Sep. 2015. [Online]. Available: <http://arxiv.org/abs/1409.0544>

- [86] H. Y. Chen and D. E. Holz, “Gamma-ray-burst beaming and gravitational-wave observations,” *Physical Review Letters*, vol. 111, 2013. [Online]. Available: <http://arxiv.org/abs/1206.0703>
- [87] N. R. Tanvir, A. J. Levan, A. S. Fruchter, J. Hjorth, R. A. Hounsell, K. Wiersema, and R. L. Tunnicliffe, “A ‘kilonova’ associated with the short-duration  $\gamma$ -ray burst GRB 130603B,” *Nature*, vol. 500, pp. 547–9, Aug. 2013. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/23912055>
- [88] S. R. Taylor and J. R. Gair, “Cosmology with the lights off: Standard sirens in the Einstein Telescope era,” *Physical Review D*, vol. 86, no. 2, p. 24, Apr. 2012. [Online]. Available: <http://prd.aps.org/abstract/PRD/v86/i2/e023502>
- [89] C. Messenger, K. Takami, S. Gossan, L. Rezzolla, and B. S. Sathyaprakash, “Host redshifts from gravitational-wave observations of binary neutron star mergers,” *Physical Review X*, vol. 4, p. 041004, 2014. [Online]. Available: <http://arxiv.org/pdf/1312.1862>
- [90] S. Konstantinidis, P. Amaro-Seoane, and K. D. Kokkotas, “Investigating the retention of intermediate-mass black holes in star clusters using N-body simulations,” *Astronomy & Astrophysics*, vol. 557, p. A135, Aug. 2013. [Online]. Available: <http://arxiv.org/abs/1108.5175>
- [91] B. Kocsis, A. Ray, and S. Portegies Zwart, “MAPPING THE GALACTIC CENTER WITH GRAVITATIONAL WAVE MEASUREMENTS USING PULSAR TIMING,” *The Astrophysical Journal*, vol. 752, no. 1, p. 67, Oct. 2012. [Online]. Available: <http://iopscience.iop.org/0004-637X/752/1/67>
- [92] C. L. Rodriguez, I. Mandel, and J. R. Gair, “Verifying the no-hair property of massive compact objects with intermediate-mass-ratio inspirals in advanced gravitational-wave detectors,” *Physical Review D*, vol. 85, no. 6, p. 062002, Mar. 2012. [Online]. Available: <http://prd.aps.org/abstract/PRD/v85/i6/e062002>
- [93] B. Kiziltan, A. Kottas, M. De Yoreo, and S. E. Thorsett, “THE NEUTRON STAR MASS DISTRIBUTION,” *The Astrophysical Journal*, vol. 778, no. 1, p. 66, Nov. 2013. [Online]. Available: <http://stacks.iop.org/0004-637X/778/i=1/a=66?key=crossref.a2a145a1f3d44a3e1f35eb5bcddf12a1>
- [94] A. P. Konstantin and L. Yungelson, “The Evolution of Compact Binary Star Systems,” *Living Reviews in Relativity*, vol. 9, no. lrr-2006-6, 2006. [Online]. Available: <http://relativity.livingreviews.org/Articles/lrr-2006-6/>

- [95] M. Dominik, K. Belczynski, C. Fryer, D. E. Holz, E. Berti, T. Bulik, I. Mandel, and R. O’Shaughnessy, “Double Compact Objects. I. the Significance of the Common Envelope on Merger Rates,” *The Astrophysical Journal*, vol. 759, p. 52, 2012.
- [96] C. Palomba *et al.*, “Searches for continuous gravitational wave signals and stochastic backgrounds in LIGO and Virgo data,” *arXiv preprint arXiv:1201.3176*, p. 9, Jan. 2012. [Online]. Available: <http://arxiv.org/abs/1201.3176>
- [97] M. van der Sluys, “Gravitational waves from compact binaries,” *arXiv preprint arXiv:1108.1307*, p. 15, Aug. 2011. [Online]. Available: <http://arxiv.org/abs/1108.1307>
- [98] K. Riles, “Gravitational waves: Sources, detectors and searches,” *Progress in Particle and Nuclear Physics*, vol. 68, pp. 1–54, Jan. 2013. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0146641012001093>
- [99] T. Regimbau, T. Dent, W. Del Pozzo, S. Giampanis, T. G. F. Li, C. Robinson, C. Van Den Broeck, D. Meacher, C. Rodriguez, B. S. Sathyaprakash, and K. Wójcik, “Mock data challenge for the Einstein Gravitational-Wave Telescope,” *Physical Review D*, vol. 86, no. 12, p. 122001, Dec. 2012. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.86.122001>
- [100] X.-J. Zhu, E. Howell, T. Regimbau, D. Blair, and Z.-H. Zhu, “STOCHASTIC GRAVITATIONAL WAVE BACKGROUND FROM COALESCING BINARY BLACK HOLES,” *The Astrophysical Journal*, vol. 739, no. 2, p. 86, Oct. 2011. [Online]. Available: <http://iopscience.iop.org/0004-637X/739/2/86>
- [101] N. J. Cornish and T. B. Littenberg, “BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches,” *arXiv preprint arXiv: 1410.3835*, p. 34, Oct. 2014. [Online]. Available: <http://arxiv.org/abs/1410.3835>
- [102] C. D. Ott, E. P. O’Connor, S. Gossan, E. Abdikamalov, U. C. T. Gamma, and S. Drasco, “Core-Collapse supernovae, neutrinos, and gravitational waves,” *Nuclear Physics B - Proceedings Supplements*, vol. 235-236, pp. 381–387, 2013.
- [103] X. Fan, C. Messenger, and I. S. Heng, “A BAYESIAN APPROACH TO MULTI-MESSENGER ASTRONOMY: IDENTIFICATION OF GRAVITATIONAL-WAVE HOST GALAXIES,” *The Astrophysical Journal*, vol. 795, no. 1, p. 43, Oct. 2014.
- [104] R. I. Santiago Prieto, “Transient gravitational waves at r-mode frequencies from neutron stars,” Ph.D. dissertation, University of Glasgow, 2014. [Online]. Available: <http://theses.gla.ac.uk/5530/>

- [105] B. S. Sathyaprakash, “Gravitational Radiation - Observing the Dark and Dense Universe,” in *The 28th International Cosmic Ray Conference*, 2004, p. 24. [Online]. Available: <http://arxiv.org/abs/gr-qc/0405136>
- [106] J. Aasi *et al.*, “Search for gravitational waves associated with  $\gamma$ -ray bursts detected by the interplanetary network,” *Physical Review Letters*, vol. 113, pp. 1–15, 2014.
- [107] C. Capano, ”in preparation”.
- [108] J. Abadie *et al.*, “First low-latency LIGO+Virgo search for binary inspirals and their electromagnetic counterparts,” *Astronomy & Astrophysics*, vol. 541, p. 155, Dec. 2012. [Online]. Available: <http://dx.doi.org/10.1051/0004-6361/201218860>
- [109] S. Babak, R. Biswas, P. R. Brady, D. A. Brown, K. Cannon, C. D. Capano, J. H. Clayton, T. Cokelaer, J. D. E. Creighton, T. Dent, A. Dietz, S. Fairhurst, N. Fotopoulos, G. González, C. Hanna, I. W. Harry, G. Jones, D. Keppel, D. J. A. McKechnan, L. Pekowsky, S. Privitera, C. Robinson, A. C. Rodriguez, B. S. Sathyaprakash, A. S. Sengupta, M. Vallisneri, R. Vaulin, and A. J. Weinstein, “Searching for gravitational waves from binary coalescence,” *Physical Review D*, vol. 87, no. 2, p. 024033, Jan. 2013. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.87.024033>
- [110] C. Pankow, P. Brady, E. Ochsner, and R. O. Shaughnessy, “A novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences,” *arXiv preprint arXiv:1502.04370*, 2015.
- [111] B. Farr, V. Kalogera, and E. Lijten, “A more efficient approach to parallel-tempered Markov-chain Monte Carlo for the highly structured posteriors of gravitational-wave signals,” *Physical Review D*, vol. 90, no. 2, p. 024014, Jul. 2014. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.90.024014>
- [112] G. Rosi, F. Sorrentino *et al.*, “Precision measurement of the Newtonian gravitational constant using cold atoms,” *Nature*, p. 510, 518-521. [Online]. Available: <http://dx.doi.org/10.1038/nature13433>
- [113] M. Agathos, W. Del Pozzo, T. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, “TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries,” *Physical Review D*, vol. 89, no. 8, p. 082001, Apr. 2014. [Online]. Available: <http://journals.aps.org/prd/abstract/10.1103/PhysRevD.89.082001>

- [114] E. S. Agency, “Planck:The Scientific Programme,” *arXiv preprint arXiv:astro-ph/0604069*, pp. 1–152, 2006. [Online]. Available: <http://www.informaworld.com/index/905664920.pdf>
- [115] M. Farhang, J. R. Bond, O. Doré, and C. B. Netterfield, “Primordial Gravitational Wave Detectability with Deep Small-Sky CMB Experiments,” *Astrophysical Journal*, vol. 771, no. 1, p. 12, Aug. 2013. [Online]. Available: <http://arxiv.org/abs/1108.2043>
- [116] L. Book, M. Kamionkowski, and F. Schmidt, “Lensing of 21-cm Fluctuations by Primordial Gravitational Waves,” *Physical Review Letters*, vol. 108, no. 21, p. 211301, May 2012. [Online]. Available: <http://prl.aps.org/abstract/PRL/v108/i21/e211301>
- [117] M. Colpi, S. Callegari, M. Dotti, and L. Mayer, “Massive black hole binary evolution in gas-rich mergers,” *Classical and Quantum Gravity*, vol. 26, no. 9, p. 8, May 2009.
- [118] M. M. Kasliwal and S. Nissanke, “ON DISCOVERING ELECTROMAGNETIC EMISSION FROM NEUTRON STAR MERGERS: THE EARLY YEARS OF TWO GRAVITATIONAL WAVE DETECTORS,” *The Astrophysical Journal*, vol. 789, no. 1, p. L5, Jul. 2014. [Online]. Available: <http://stacks.iop.org/2041-8205/789/i=1/a=L5?key=crossref.c67ffe2830df55057867eff48d056a73>
- [119] J. Centrella, S. Nissanke, and R. Williams, “Gravitational Waves and Time-Domain Astronomy,” *Proceedings of the International Astronomical Union*, vol. 7, no. S285, pp. 191–198, Apr. 2012. [Online]. Available: <http://www.journals.cambridge.org/abstract.S1743921312000592>
- [120] V. Predoi and K. Hurley, “Search for gravitational waves associated with the InterPlanetary Network short gamma ray bursts,” *Journal of Physics: Conference Series*, vol. 363, p. 012034, Dec. 2012. [Online]. Available: <http://stacks.iop.org/1742-6596/363/i=1/a=012034?key=crossref.eae8755a51233eb078c230be979f0cd5>
- [121] K. Postnov and L. Yungelson, “Close Binary White Dwarfs, Neutron Stars, Black Holes: Formation, Evolution, and Related Phenomena,” *Arxiv preprint arXiv:1403.4754*, p. 181, Mar. 2014. [Online]. Available: <http://arxiv.org/abs/1403.4754>
- [122] X.-F. Wu, H. Gao, X. Ding, B. Zhang, Z.-G. Dai, and J.-Y. Wei, “A DOUBLE NEUTRON STAR MERGER ORIGIN FOR THE COSMOLOGICAL RELATIVISTIC FADING SOURCE PTF11agg?” *The Astrophysical Journal*, vol. 781, no. 1, p. L10, Jan. 2014. [Online]. Available: <http://iopscience.iop.org/2041-8205/781/1/L10>



- [123] I. Bartos, A. P. S. Crotts, and M. Szabolcs, “Galaxy Survey On The Fly : Prospects of Rapid Galaxy Cataloging to Aid the Electromagnetic Follow-up of Gravitational-wave Observations,” *arXiv preprint arXiv: 1410.0677*, pp. 1–4, 2014.
- [124] I. Bartos and T. L. Huard, “James Webb Space Telescope can Detect Kilonovae in Gravitational Wave Follow-up,” *in prep.*, 2015.
- [125] L. Singer, “The needle in the hundred square degree haystack: The hunt for binary neutron star mergers with LIGO and Palomar Transient Factory,” Ph.D. dissertation, California Institute of Technology, 2015. [Online]. Available: <http://thesis.library.caltech.edu/8739/>
- [126] B. S. Sathyaprakash, B. Schutz, and C. V. D. Broeck, “Cosmography with the Einstein Telescope,” *Classical and Quantum Gravity*, vol. 27, no. 21, p. 215006, 2009. [Online]. Available: <http://arxiv.org/abs/0906.4151>
- [127] N. L. Christensen, f. t. L. S. Collaboration, and t. V. Collaboration, “Multimessenger Astronomy,” *arXiv preprint arXiv: 1105.5843*, May 2011. [Online]. Available: <http://arxiv.org/abs/1105.5843>
- [128] A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry, “Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant,” *The Astronomical Journal*, vol. 116, pp. 1009–1038, Sep. 1998.
- [129] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, J. C. Lee, N. J. Nunes, R. Pain, C. R. Pennypacker, R. Quimby, C. Lidman, R. S. Ellis, M. Irwin, R. G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B. J. Boyle, A. V. Filippenko, T. Matheson, A. S. Fruchter, N. Panagia, H. J. M. Newberg, W. J. Couch, and T. S. C. Project, “Measurements of  $\Omega$  and  $\Lambda$  from 42 High-Redshift Supernovae,” *The Astrophysical Journal*, vol. 517, pp. 565–586, Jun. 1999.
- [130] A. G. Riess, L. Macri, S. Casertano, H. Lampeitl, H. C. Ferguson, A. V. Filippenko, S. W. Jha, W. Li, and R. Chornock, “A 3% Solution: Determination of the Hubble Constant with the Hubble Space Telescope and Wide Field Camera 3,” *The Astrophysical Journal*, vol. 730, p. 119, Apr. 2011.
- [131] A. Nishizawa, K. Yagi, A. Taruya, and T. Tanaka, “Gravitational-wave standard siren without redshift identification,” in *Journal of Physics: Conference Series*, vol. 363, 2012, p. 012052.

- [132] W. Del Pozzo, “Inference of cosmological parameters from gravitational waves: Applications to second generation interferometers,” *Physical Review D*, vol. 86, no. 4, p. 043011, Aug. 2012. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.86.043011>
- [133] M. Biesiada, X. Ding, A. Piorkowska, and Z.-H. Zhu, “Strong gravitational lensing of gravitational waves from double compact binaries - perspectives for the Einstein Telescope,” *Arxiv preprint arXiv:1409.8360*, Sep. 2014. [Online]. Available: <http://arxiv.org/abs/1409.8360v1>
- [134] M. Sereno, P. Jetzer, A. Sesana, and M. Volonteri, “Cosmography with strong lensing of LISA gravitational wave sources,” *Monthly Notices of the Royal Astronomical Society*, vol. 415, no. April, pp. 2773–2781, 2011.
- [135] S. Yu and C. S. Jeffery, “The influence of star formation history on the gravitational wave signal from close double degenerates in the thin disc,” *Monthly Notices of the Royal Astronomical Society*, vol. 429, no. November, pp. 1602–1608, 2013.
- [136] P. Raffai, L. Gondán, I. S. Heng, N. Kelecsényi, J. Logue, Z. Márka, and S. Márka, “Optimal networks of future gravitational-wave telescopes,” *Classical and Quantum Gravity*, vol. 30, no. 15, p. 155004, Aug. 2013. [Online]. Available: <http://stacks.iop.org/0264-9381/30/i=15/a=155004?key=crossref.ece86b0eefe83f795ae2d006297ebe2e>
- [137] B. Schutz, “Networks of gravitational wave detectors and three figures of merit,” *Classical and Quantum Gravity*, vol. 28, no. 12, p. 125023, Jun. 2011. [Online]. Available: <http://stacks.iop.org/0264-9381/28/i=12/a=125023?key=crossref.f15bd0947108bf487a9971d94688381c>
- [138] I. W. Harry and S. Fairhurst, “Targeted coherent search for gravitational waves from compact binary coalescences,” *Physical Review D*, vol. 83, no. 8, p. 084002, Apr. 2011. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.83.084002>
- [139] T. Futamase and Y. Itoh, “The Post-Newtonian Approximation for Relativistic Compact Binaries,” *Living Reviews in Relativity*, vol. 10, 2007. [Online]. Available: <http://www.livingreviews.org/lrr-2007-2>
- [140] M. Boyle, “Accurate gravitational waveforms from binary black-hole systems,” Ph.D. dissertation, California Institute of Technology, 2009. [Online]. Available: <http://thesis.library.caltech.edu/143/>
- [141] H. Tagoshi, C. K. Mishra, A. Pai, and K. G. Arun, “Parameter estimation of neutron star-black hole binaries using an advanced gravitational-wave detector network:

- Effects of the full post-Newtonian waveform,” *Physical Review D*, vol. 90, no. 2, p. 024053, Jul. 2014. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.90.024053>
- [142] E. Barausse and L. Lehner, “Post-Newtonian approach to black hole-fluid systems,” *Physical Review D*, vol. 88, no. 2, p. 024029, Jul. 2013. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.88.024029>
- [143] B. Allen, “A chi-squared time-frequency discriminator for gravitational wave detection,” *Physical Review D*, vol. 71, no. 6, p. 18, May 2004. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.71.062001>
- [144] L. Blanchet, B. R. Iyer, C. M. Will, and A. G. Wiseman, “Gravitational waveforms from inspiralling compact binaries to second-post-Newtonian order,” *Classical and Quantum Gravity*, vol. 13, no. 4, pp. 575–584, Apr. 1996. [Online]. Available: <http://stacks.iop.org/0264-9381/13/i=4/a=002?key=crossref.5c9444e46619681b5ed2702080e666f3>
- [145] J. Hartung and J. Steinhoff, “Next-to-next-to-leading order post-Newtonian spin-orbit Hamiltonian for self-gravitating binaries,” *Annalen der Physik*, vol. 523, no. 10, pp. 783–790, Oct. 2011. [Online]. Available: <http://doi.wiley.com/10.1002/andp.201100094>
- [146] F. Biscani and S. Carloni, “A first-order secular theory for the post-Newtonian two-body problem with spin - I. The restricted case,” *Monthly Notices of the Royal Astronomical Society*, vol. 428, no. 3, pp. 2295–2310, Nov. 2012. [Online]. Available: <http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/sts198>
- [147] K. Cannon, R. Cariou, A. Chapman, M. Crispin-Ortuzar, N. Fotopoulos, M. Frei, C. Hanna, E. Kara, D. Keppel, L. Liao, S. Privitera, A. Searle, L. Singer, and A. Weinstein, “Toward Early-Warning Detection of Gravitational Waves From Compact Binary Coalescence,” *The Astrophysical Journal*, vol. 748, no. 2, p. 136, Apr. 2012. [Online]. Available: <http://stacks.iop.org/0004-637X/748/i=2/a=136?key=crossref.776ef7a7b765fd0ef6d757fd3356a61a>
- [148] S. Hooper, S. Chung, J. Luan, D. Blair, Y. Chen, and L. Wen, “Summed parallel infinite impulse response filters for low-latency detection of chirping gravitational waves,” *Physical Review D*, vol. 86, no. 2, p. 024012, Jul. 2012. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.86.024012>
- [149] J. Abadie, B. Abbott, R. Abbott, and T. Abbott, “Search for gravitational waves from low mass compact binary coalescence in LIGO’s sixth science run and Virgo’s

- science runs 2 and 3,” *Physical Review D*, vol. 85, no. 8, p. 082002, Apr. 2012. [Online]. Available: <http://prd.aps.org/abstract/PRD/v85/i8/e082002>
- [150] T. Dent and J. Veitch, “Optimizing gravitational-wave searches for a population of coalescing binaries: Intrinsic parameters,” *Physical Review D*, vol. 89, no. 6, p. 062002, Mar. 2014. [Online]. Available: <http://journals.aps.org/prd/abstract/10.1103/PhysRevD.89.062002>
- [151] C. Chatfield, *The analysis of time series: an introduction*, 6th ed. Florida, US: CRC Press, 2004.
- [152] C. L. Rodriguez, B. Farr, W. M. Farr, and I. Mandel, “Inadequacies of the Fisher Information Matrix in gravitational-wave parameter estimation,” *Physical Review D*, vol. 88, no. 8, p. 084013, Aug. 2013. [Online]. Available: <http://prd.aps.org/abstract/PRD/v88/i8/e084013>
- [153] T. J. Loredo, “On the future of astrostatistics: statistical foundations and statistical practice,” *Statistical Challenges in Modern Astronomy V*, pp. 303–308, Aug. 2012. [Online]. Available: <http://arxiv.org/abs/1208.3035>
- [154] —, “Bayesian astrostatistics: a backward look to the future,” *Astrostatistical Challenges for the New Astronomy*, vol. c, no. August 2011, Aug. 2012. [Online]. Available: <http://arxiv.org/abs/1208.3036>
- [155] J. VanderPlas, “Frequentism and Bayesianism: A Python-driven Primer,” in *THE 13th PYTHON IN SCIENCE CONF.*, S. van der Walt and J. Bergstra, Eds., no. Scipy, Nov. 2014, pp. 90–98. [Online]. Available: <http://arxiv.org/abs/1411.5018v1>
- [156] P. Gregory, *Bayesian logical data analysis for the physical sciences*. Cambridge: Cambridge University Press, 2005. [Online]. Available: <http://ebooks.cambridge.org/ref/id/CBO9780511791277>
- [157] D. S. Sivia and J. Skilling, *Data analysis. A Bayesian tutorial. 2nd ed.* Oxford: Oxford University Press., 2006.
- [158] D. Spiegelhalter, “Court of appeal bans bayesian probability (and sherlock holmes),” 2013. [Online]. Available: <http://understandinguncertainty.org/court-appeal-bans-bayesian-probability-and-sherlock-holmes>
- [159] M. Kilbinger, D. Wraith, C. P. Robert, K. Benabed, O. Cappé, J.-F. Cardoso, G. Fort, S. Prunet, and F. R. Bouchet, “Bayesian model comparison in cosmology with Population Monte Carlo,” *Monthly Notices of the Royal Astronomical Society*, vol. 2390, pp. 2381–2390, Apr. 2010. [Online]. Available: <http://doi.wiley.com/10.1111/j.1365-2966.2010.16605.x>

- [160] P. Green, “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995. [Online]. Available: <http://biomet.oxfordjournals.org/content/82/4/711.short>
- [161] Y.-M. Hu, M. Pitkin, I. Heng, and M. Hendry, “Glitch or anti-glitch: a Bayesian view,” *Astrophysical Journal Letters*, vol. 784, no. 2, p. L41, Apr. 2014. [Online]. Available: <http://stacks.iop.org/2041-8205/784/i=2/a=L41?key=crossref.596227231ee33f97fe86a90f43a9be65>
- [162] R. F. Archibald, V. M. Kaspi, C. Y. Ng, K. N. Gourgouliatos, D. Tsang, P. Scholz, A. P. Beardmore, N. Gehrels, and J. A. Kennea, “An Anti-Glitch in a Magnetar,” *Nature*, vol. 497, no. 7451, pp. 591–593, May 2013.
- [163] H. Jeffreys, *Theory of Probability*, 3rd ed. Oxford, England: Oxford, 1961.
- [164] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. [Online]. Available: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Information+Theory+,+Inference+,+and+Learning+Algorithms#8>
- [165] E. Cameron and A. Pettitt, “Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis,” *Statistical Science*, vol. 29, no. 3, pp. 397–419, Aug. 2014. [Online]. Available: <http://projecteuclid.org/euclid.ss/1411437520>
- [166] J. Skilling, “Nested sampling,” *AIP Conference Proceedings*, vol. 735, no. 1, pp. 395–405, 2004. [Online]. Available: <http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.1835238>
- [167] —, “Nested sampling for general Bayesian computation,” *Bayesian Analysis*, vol. 1, no. 4, pp. 833–859, Dec. 2006. [Online]. Available: <http://projecteuclid.org/euclid.ba/1340370944>
- [168] F. Feroz, M. P. Hobson, and M. Bridges, “MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics,” *Monthly Notices of the Royal Astronomical Society*, vol. 398, no. 4, pp. 1601–1614, Oct. 2009. [Online]. Available: <http://mnras.oxfordjournals.org/cgi/doi/10.1111/j.1365-2966.2009.14548.x>
- [169] W. M. Farr, I. Mandel, and D. Stevens, “An Efficient Interpolation Technique for Jump Proposals in Reversible-Jump Markov Chain Monte Carlo Calculations,” *arXiv preprint arXiv: 1104.0984*, Apr. 2011. [Online]. Available: <http://arxiv.org/abs/1104.0984>

- [170] N. Karnesis, M. Nofrarias, C. F. Sopuerta, F. Gibert, M. Armano, H. Audley, G. Congedo, I. Diepholz, L. Ferraioli, M. Hewitson, M. Hueller, N. Korsakova, P. W. McNamara, E. Plagnol, and S. Vitale, “Bayesian model selection for LISA pathfinder,” *Physical Review D*, vol. 89, no. 6, p. 062001, Mar. 2014. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.89.062001>
- [171] L. Lentati, P. Alexander, M. P. Hobson, F. Feroz, R. van Haasteren, K. J. Lee, and R. M. Shannon, “TEMPONEST: a Bayesian approach to pulsar timing analysis,” *Monthly Notices of the Royal Astronomical Society*, vol. 437, no. 3, pp. 3004–3023, Nov. 2013. [Online]. Available: <http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/stt2122>
- [172] R. Nuzzo, “Scientific method: statistical errors,” *Nature*, vol. 506, no. 7487, pp. 150–152, Feb. 2014. [Online]. Available: <http://dx.doi.org/10.1038/506150a>
- [173] K. Mikkelsen, S. K. Næss, and H. K. Eriksen, “Grid-Based Exploration of Cosmological Parameter Space With Snake,” *The Astrophysical Journal*, vol. 777, no. 2, p. 172, Nov. 2013. [Online]. Available: <http://stacks.iop.org/0004-637X/777/i=2/a=172?key=crossref.e523e009acc1b0e0bc6b89c62c9a43d6>
- [174] R. Andrae, “Error estimation in astronomy: A guide,” *arXiv preprint arXiv:1009.2755*, p. 23, 2010. [Online]. Available: <http://arxiv.org/pdf/1009.2755>
- [175] M. Bastarrika, “Bayesian analysis of burst gravitational waves from galactic neutron stars,” Ph.D. dissertation, University of Glasgow, 2010. [Online]. Available: <http://theses.gla.ac.uk/2363/>
- [176] R. Allison and J. Dunkley, “Comparison of sampling techniques for Bayesian parameter estimation,” *Monthly Notices of the Royal Astronomical Society*, vol. 437, no. 4, pp. 3918–3928, Dec. 2013. [Online]. Available: <http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/stt2190>
- [177] C. Röver, “Bayesian inference on astrophysical binary inspirals based on gravitational-wave measurements,” Ph.D. dissertation, University of Auckland, 2007. [Online]. Available: <https://researchspace.auckland.ac.nz/handle/2292/2356>
- [178] H. Haario, E. Saksman, and J. Tamminen, “An adaptive Metropolis algorithm,” *Bernoulli*, vol. 7, no. 2, pp. 223–242, 2001. [Online]. Available: <http://www.jstor.org/stable/10.2307/3318737>
- [179] P. Green, “Bayesian System Identification of Nonlinear Dynamical Systems using a Fast MCMC Algorithm,” *Proceedings of ENOC 2014, European Nonlinear*

- Dynamics Conference. ENOC 2014, European Nonlinear Dynamics Conference*, no. 3, 2014. [Online]. Available: <http://eprints.whiterose.ac.uk/81829/>
- [180] L. Tierney and a. Mira, “Some adaptive monte carlo methods for Bayesian inference.” *Statistics in medicine*, vol. 18, no. 17-18, pp. 2507–15, 1999. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/10474156>
- [181] F. Feroz and J. Skilling, “Exploring multi-modal distributions with nested sampling,” in *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, no. 2, Dec. 2013, pp. 106–113. [Online]. Available: <http://link.aip.org/link/APCPCS/v1553/i1/p106/s1&Agg=doi>
- [182] S. Martiniani, J. D. Stevenson, D. J. Wales, and D. Frenkel, “Superposition Enhanced Nested Sampling,” *Physical Review X*, vol. 4, no. 3, p. 031034, Feb. 2014. [Online]. Available: <http://arxiv.org/abs/1402.6306><http://link.aps.org/doi/10.1103/PhysRevX.4.031034>
- [183] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.
- [184] W. K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.
- [185] A. Gelman and D. Rubin, “Inference from iterative simulation using multiple sequences,” *Statistical science*, vol. 7, no. 4, p. 457, 1992. [Online]. Available: <http://www.jstor.org/stable/10.2307/2246093>
- [186] L. Tierney, “Markov Chains for Exploring Posterior Distributions,” *the Annals of Statistics*, vol. 22, no. 4, pp. 1701–1728, 1994. [Online]. Available: <http://www.jstor.org/stable/2242477>
- [187] A. Peraiah, *An Introduction to Radiative Transfer: Methods and Applications in Astrophysics*. Cambridge: Cambridge University Press, 2001. [Online]. Available: <http://www.cambridge.org/gb/academic/subjects/astronomy/astrophysics/introduction-radiative-transfer-methods-and-applications-astrophysics>
- [188] W. Kim and K. M. Lee, “Scanline sampler without detailed balance: An efficient mcmc for mrf optimization,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR ’14. Washington,

- DC, USA: IEEE Computer Society, 2014, pp. 1354–1361. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2014.176>
- [189] M. D. Weinberg, “Computing the Bayes Factor from a Markov Chain Monte Carlo Simulation of the Posterior Distribution,” *Bayesian Analysis*, vol. 7, no. 3, pp. 737–770, Sep. 2012. [Online]. Available: <http://projecteuclid.org/euclid.ba/1346158782>
- [190] K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, and B. Placek, “Bayesian Evidence and Model Selection,” *arXiv preprint arXiv: ...*, p. 39, Nov. 2014. [Online]. Available: <http://arxiv.org/abs/1411.3013>
- [191] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, “Importance Nested Sampling and the MultiNest Algorithm,” *arXiv preprint arXiv: 1306.2144*, p. 28, 2013. [Online]. Available: <http://arxiv.org/abs/1306.2144>
- [192] R. H. Swendsen and J. S. Wang, “Replica Monte Carlo Simulation of Spin-Glasses,” *Physical Review Letters*, vol. 57, no. 21, pp. 2607–2609, Nov. 1986.
- [193] R. Neal, “Sampling from multimodal distributions using tempered transitions,” *Statistics and computing*, vol. 6, pp. 353–366, 1996. [Online]. Available: <http://www.springerlink.com/index/X6417MM4H7845735.pdf>
- [194] D. J. Earl and M. W. Deem, “Parallel tempering: Theory, applications, and new perspectives,” *Physical Chemistry Chemical Physics*, vol. 7, no. 23, p. 3910, Aug. 2005. [Online]. Available: <http://arxiv.org/abs/physics/0508111http://xlink.rsc.org/?DOI=b509983h>
- [195] W. D. Voudsen, W. M. Farr, and I. Mandel, “Dynamic temperature selection for parallel-tempering in Markov chain Monte Carlo simulations,” *in prep.*, 2015.
- [196] A. Mira, “On Metropolis-Hastings algorithms with delayed rejection,” *Metron*, 2001. [Online]. Available: [http://eco.uninsubria.it/Dipeco/Quaderni/files/QF2000\\_6\\_.pdf](http://eco.uninsubria.it/Dipeco/Quaderni/files/QF2000_6_.pdf)
- [197] P. Green and A. Mira, “Delayed rejection in reversible jump Metropolis-Hastings,” *Biometrika*, vol. 88, no. 4, pp. 1035–1053, 2001. [Online]. Available: <http://www.jstor.org/stable/10.2307/2673700>
- [198] M. Trias, A. Vecchio, and J. Veitch, “Delayed rejection schemes for efficient Markov-Chain Monte-Carlo sampling of multimodal distributions,” *arXiv preprint arXiv:0904.2207*, p. 32, Apr. 2009. [Online]. Available: <http://arxiv.org/abs/0904.2207>



- [199] P. H. Peskun, "Optimum Monte-Carlo Sampling Using Markov Chains," *Biometrika*, vol. 60, no. 3, p. 607, Dec. 1973. [Online]. Available: <http://www.jstor.org/stable/2335011?origin=crossref>
- [200] P. C. Gregory, K. H. Knuth, A. Caticha, J. L. Center, A. Giffin, and C. C. Rodriguez, "A Bayesian re-analysis of HD 11964: evidence for three planets," *AIP Conference Proceedings*, vol. 954, no. 1, pp. 307–314, 2007. [Online]. Available: <http://link.aip.org/link/APCPCS/v954/i1/p307/s1&Agg=doi>
- [201] J. Goodman and J. Weare, "Ensemble samplers with affine invariance," *Communications in Applied Mathematics and Computational Science*, vol. 5, no. 1, pp. 65–80, 2010.
- [202] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "emcee : The MCMC Hammer," *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 925, pp. 306–312, Mar. 2013. [Online]. Available: <http://www.jstor.org/stable/10.1086/670067>
- [203] J. Veitch and A. Vecchio, "Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network," *Physical Review D*, vol. 81, no. 6, p. 062003, Mar. 2010. [Online]. Available: <http://prd.aps.org/abstract/PRD/v81/i6/e062003>
- [204] B. Link, R. I. Epstein, and K. A. van Riper, "Pulsar glitches as probes of neutron star interiors," *Nature*, vol. 359, pp. 616–618, Oct. 1992. [Online]. Available: <http://www.nature.com/nature/journal/v359/n6396/abs/359616a0.html>
- [205] C. Thompson, R. C. Duncan, P. M. Woods, C. Kouveliotou, M. H. Finger, and J. van Paradijs, "Physical Mechanisms for the Variable Spindown and Light Curve of SGR 1900+14," *The Astronomical Journal*, vol. 543, no. 1, pp. 340–350, Nov. 2000.
- [206] J. a. Pons and N. Rea, "Modeling Magnetar Outbursts: Flux Enhancements and the Connection With Short Bursts and Glitches," *The Astrophysical Journal*, vol. 750, no. 1, p. L6, May 2012. [Online]. Available: <http://stacks.iop.org/2041-8205/750/i=1/a=L6?key=crossref.5d602b029bf5e0d8d56c3813ca0e1b7b>
- [207] H. Tong, "Anti-glitch of magnetar 1E 2259+586 in the wind braking scenario," *The Astronomical Journal*, vol. 784, no. 2, p. 86, Jun. 2014. [Online]. Available: <http://iopscience.iop.org/0004-637X/784/2/86/>
- [208] M. Lyutikov, "Magnetospheric "anti-glitches" in magnetars," *arXiv:1306.2264*, Jun. 2013.

- [209] J. I. Katz, “Implications of an anti-glitch in AXP/SGR,” *Astrophysics and Space Science*, Dec. 2013.
- [210] Y. F. Huang and J. J. Geng, “ANTI-GLITCH INDUCED BY COLLISION OF A SOLID BODY WITH THE MAGNETAR 1E 2259+586,” *The Astrophysical Journal*, vol. 782, p. L20, Oct. 2014. [Online]. Available: <http://adsabs.harvard.edu/abs/2014ApJ...782L..20H>
- [211] R. Ouyed, D. Leahy, and N. Koning, ““Anti-glitches” in the Quark-Nova model for AXPs I,” pp. 715–718, Jul. 2014. [Online]. Available: <http://arxiv.org/abs/1307.1386>
- [212] D. Lorimer and M. Kramer, *Handbook of Pulsar Astronomy*, 1st ed., ser. Cambridge Observing Handbooks for Research Astronomers. Cambridge, U.K.; New York, U.S.A: Cambridge University Press, 2005, vol. 4.
- [213] G. B. Hobbs, R. T. Edwards, and R. N. Manchester, “tempo2, a new pulsar-timing package - I. An overview,” *MNRAS*, vol. 369, no. 2, pp. 655–672, Jun. 2006.
- [214] R. Dib, “RXTE monitoring of 5 anomalous x-ray pulsars,” Ph.D. dissertation, McGill University, Canada, Sep. 2009.
- [215] V. Kaspi and F. Gavriil, “A major soft gamma repeater-like outburst and rotation glitch in the no-longer-so-anomalous X-ray pulsar 1E 2259+ 586,” *The Astronomical Journal*, vol. 588, no. 2, pp. L93–L96, May 2003.
- [216] B. Link and R. I. Epstein, “Thermally Driven Neutron Star Glitches,” *The Astronomical Journal*, vol. 457, p. 844, Feb. 1996.
- [217] C. Thompson and R. C. Duncan, “The soft gamma repeaters as very strongly magnetized neutron stars - I. Radiative mechanism for outbursts,” *Monthly Notices of the Royal Astronomical Society*, vol. 275, pp. 255–300, Jul. 1995.
- [218] R. C. Duncan, “Astrophysics: A glimpse inside a magnetar.” *Nature*, vol. 497, no. 7451, pp. 574–6, May 2013.
- [219] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *SCIENCE*, vol. 220, no. 4598, pp. 671–680, 1983.
- [220] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. New York, NY, USA: Cambridge University Press, 2007.
- [221] J. Crowder, N. Cornish, and J. Reddinger, “LISA data analysis using genetic algorithms,” *Physical Review D*, vol. 73, no. 6, p. 063011, Mar. 2006. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.73.063011>

- [222] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.
- [223] R. M. Neal, "MCMC using Hamiltonian dynamics," *ArXiv e-prints*, Jun. 2012. [Online]. Available: <http://arxiv.org/abs/1206.1901>
- [224] J. Abadie *et al.*, "Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors," *Classical and Quantum Gravity*, vol. 27, no. 17, p. 173001, Sep. 2010. [Online]. Available: <http://stacks.iop.org/0264-9381/27/i=17/a=173001?key=crossref.8cd38004165a7603a0d966b5b7b8f15d>
- [225] the ET science team. (2011) Einstein gravitational wave telescope conceptual design study. [Online]. Available: <https://tds.ego-gw.it/ql/?c=7954>
- [226] B. S. Sathyaprakash *et al.*, "Scientific objectives of Einstein Telescope," *Classical and Quantum Gravity*, vol. 29, no. 12, p. 124013, Jun. 2012. [Online]. Available: <http://iopscience.iop.org/0264-9381/29/12/124013>
- [227] E. Chassande-Mottin, M. Hendry, P. J. Sutton, and S. Márka, "Multimessenger astronomy with the Einstein Telescope," *General Relativity and Gravitation*, vol. 43, no. 2, p. 26, Apr. 2010. [Online]. Available: <http://link.springer.com/10.1007/s10714-010-1019-z>
- [228] M. Coughlin and J. Harms, "Seismic topographic scattering in the context of GW detector site selection," *Classical and Quantum Gravity*, vol. 29, no. 7, p. 075004, Apr. 2012. [Online]. Available: <http://stacks.iop.org/0264-9381/29/i=7/a=075004?key=crossref.de58befa252c59bc146a9867c2621c8c>
- [229] J. Veitch, I. Mandel, B. Aylott, B. Farr, V. Raymond, C. Rodriguez, M. van der Sluys, V. Kalogera, and A. Vecchio, "Estimating parameters of coalescing compact binaries with proposed advanced detector networks," *Physical Review D*, vol. 85, no. 10, p. 104045, May 2012. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.85.104045>
- [230] S. Fairhurst, "Improved source localization with LIGO-India," *Journal of Physics: Conference Series*, vol. 484, p. 012007, Mar. 2014. [Online]. Available: <http://stacks.iop.org/1742-6596/484/i=1/a=012007?key=crossref.d2fa0a752cbd3c0273bff0334adee375>

- [231] ———, “Source localization with an advanced gravitational wave detector network,” *Classical and Quantum Gravity*, vol. 28, no. 10, p. 105021, 2011. [Online]. Available: <http://iopscience.iop.org/0264-9381/28/10/105021>
- [232] M. van der Sluys, V. Raymond, I. Mandel, C. Rver, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio, “Parameter estimation of spinning binary inspirals using markov chain monte carlo,” *Classical and Quantum Gravity*, vol. 25, no. 18, p. 184011, 2008. [Online]. Available: <http://stacks.iop.org/0264-9381/25/i=18/a=184011>
- [233] V. Raymond, M. V. van der Sluys, I. Mandel, V. Kalogera, C. Rver, and N. Christensen, “Degeneracies in sky localization determination from a spinning coalescing binary through gravitational wave observations: a markov-chain monte carlo analysis for two detectors,” *Classical and Quantum Gravity*, vol. 26, no. 11, p. 114007, 2009. [Online]. Available: <http://stacks.iop.org/0264-9381/26/i=11/a=114007>
- [234] C. J. F. T. Braak, “A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces,” *Statistics and Computing*, vol. 16, no. 3, pp. 239–249, Sep. 2006. [Online]. Available: <http://www.springerlink.com/index/10.1007/s11222-006-8769-1>
- [235] Y.-M. Hu, M. Hendry, and I. S. Heng, “Efficient Exploration of Multi-Modal Posterior Distributions,” *Arxiv preprint arXiv:1408.3969*, p. 6, Aug. 2014. [Online]. Available: <http://arxiv.org/abs/1408.3969>
- [236] Aster gdem worldwide elevation data map. <http://asterweb.jpl.nasa.gov/gdem.asp>.
- [237] Natural earth. <http://www.naturalearthdata.com>.
- [238] Nasa visible earth. <http://visibleearth.nasa.gov/view.php?id=55167>.
- [239] World database on protected areas. <http://protectedplanet.net/>.
- [240] *Exclusion based on tectonic plate lines.*
- [241] K. M. Grski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, “Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere,” *The Astrophysical Journal*, vol. 622, no. 2, p. 759, 2005. [Online]. Available: <http://stacks.iop.org/0004-637X/622/i=2/a=759>
- [242] S. Merkowitz and W. Johnson, “Spherical gravitational wave antennas and the truncated icosahedral arrangement.” *Physical review D: Particles and*

- fields*, vol. 51, no. 6, pp. 2546–2558, Mar. 1995. [Online]. Available: <http://journals.aps.org/prd/abstract/10.1103/PhysRevD.51.2546>
- [243] —, “Techniques for detecting gravitational waves with a spherical antenna,” *Physical Review D*, vol. 56, no. 12, pp. 7513–7528, Dec. 1997. [Online]. Available: <http://link.aps.org/doi/10.1103/PhysRevD.56.7513>
- [244] J. Aasi *et al.*, “Prospects for localization of gravitational wave transients by the Advanced LIGO and Advanced Virgo Observatories,” *arXiv preprint arXiv:1304.0670*, p. 18, Apr. 2013.
- [245] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, and et al., “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” *Physics Letters B*, vol. 716, pp. 1–29, Sep. 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S037026931200857X>
- [246] S. Vitale, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, I. Mandel, B. Aylott, and J. Veitch, “Effect of calibration errors on Bayesian parameter estimation for gravitational wave signals from inspiral binary systems in the advanced detectors era,” *Physical Review D*, vol. 85, no. 6, p. 064034, Mar. 2012. [Online]. Available: <http://prd.aps.org/abstract/PRD/v85/i6/e064034>
- [247] J. Abadie *et al.*, “Search for gravitational waves from compact binary coalescence in LIGO and Virgo data from S5 and VSR1,” *Physical Review D - Particles, Fields, Gravitation and Cosmology*, vol. 82, p. 102001, 2010. [Online]. Available: <http://journals.aps.org/prd/abstract/10.1103/PhysRevD.82.102001>
- [248] D. Keppel, “Signatures and dynamics of compact binary coalescences and a search in LIGO’s S5 data,” Ph.D. dissertation, California Institute of Technology, 2009. [Online]. Available: <http://thesis.library.caltech.edu/1901/>